NYPD Shooting Analysis

Brandon Fry

1/29/2022

Objective

The objective of this analysis is to determine how gun violence has changed over the last decade before the Covid pandemic occured. This data could provide useful insight to see if improvements were being seen in gun violence, and help to provide insight into staffing and resource requirements.

We also aim to predict what the future of gun violence looks like, as well as how 3 different cities compare and how they will compare in the future. This analysis could provide benefit to the NYPD so they could have a better understanding of which cities are best battling gun violence, and also how they may need to reallocate resources to different jurisdictions.

Process

We will use historical data from the NYPD for three different cities, which include Brooklyn, Queens, and the Bronx. We will filter data up through 2019 before Covid became widespread. This allows us to see how progress was before being impacted by the pandemic. We will then use the tools and packages available in R to tidy up the data and view only data for the cities of interest.

The next step is to group shootings into groups based on the year that it occured. By tidying the data, it allows us to remove the data not necessary for this analysis and focus only on the data that peratins to this study.

Once the data is tidy, we will perform a few visualizations. The first will be a view of the raw data plotted on a graph, which will allow us to get an initial understanding of the trends represented in the data. The second graph takes it a step further by generating a prediction model, and plotting it relative to the raw data. This allows us to validate the trend in shootings, and also validate that our prediction model has some level of accuracy. The third plot shows the prediction model carried out over the next data. This allows us to see how gun violence may change in the future, and how the different cities may become better or worst than others.

```
library("tidyverse")
library(plyr)

#Gather data from source
url_in = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
data <- read_csv(url_in)

#Parse data from Brooklyn to see gun crimes commited for each month
data_brook <- subset(data %>% filter(BORO == "BROOKLYN"), select = +c(OCCUR_DATE, OCCUR_TIME, BORO))
data_brook$OCCUR_DATE <- as.Date(data_brook$OCCUR_DATE,
                        format = "%m/%d/%Y")
data_brook$Year <- format(as.Date(data_brook$OCCUR_DATE), "%Y")
```

```r
data_brook <- count(data_brook, "Year")
data_brook <- na.omit(data_brook)
data_brook <- filter(data_brook, Year < 2020)
data_brook$Year <- as.numeric(data_brook$Year, format='%Y')
data_brook <- data_brook %>% mutate(city = "Brooklyn")

#Parse data from Queens to see gun crimes commited for each month
data_queens <- subset(data %>% filter(BORO == "QUEENS"), select = +c(OCCUR_DATE, OCCUR_TIME, BORO))
data_queens$OCCUR_DATE <- as.Date(data_queens$OCCUR_DATE,
                        format = "%m/%d/%Y")
data_queens$Year <- format(as.Date(data_queens$OCCUR_DATE), "%Y")
data_queens <- count(data_queens, "Year")
data_queens <- na.omit(data_queens)
data_queens <- filter(data_queens, Year < 2020)
data_queens$Year <- as.numeric(data_queens$Year, format='%Y')
data_queens <- data_queens %>% mutate(city = "Queens")

#Parse data from Bronx to see gun crimes commited for each month
data_bronx <- subset(data %>% filter(BORO == "BRONX"), select = +c(OCCUR_DATE, OCCUR_TIME, BORO))
data_bronx$OCCUR_DATE <- as.Date(data_bronx$OCCUR_DATE,
                        format = "%m/%d/%Y")
data_bronx$Year <- format(as.Date(data_bronx$OCCUR_DATE), "%Y")
data_bronx <- count(data_bronx, "Year")
data_bronx <- na.omit(data_bronx)
data_bronx <- filter(data_bronx, Year < 2020)
data_bronx$Year <- as.numeric(data_bronx$Year, format='%Y')
data_bronx <- data_bronx %>% mutate(city = "Bronx")
```

Data Analysis

The following plot shows the number of gun violence occurrences for each year between 2006 and 2019 for each city. From initial inspection, we can see that it appears that there is an overall decline in gun violence. We can also see that there is a clear rank between the cities, with Queens having the lowest gun violence rate, and Brooklyn having the highest gun violence rate. Another observation that can be made is that the cities have been converging to a similar level, which may indicate that a shift in rank may occur in the future

Later in this study, we will use a prediction model to help better visualize the trend and predict how it may look over the next decade, and see if our hypothesis is true.

```
ggplot(data_brook, aes(x = Year, y = freq, color=city)) +
  geom_point() +
  geom_point(data=data_queens, aes(color = city)) +
  geom_point(data=data_bronx, aes(color = city)) +
  labs(title = "Shootings Per Year", x = "Year", y = "Number of Shootings")
```
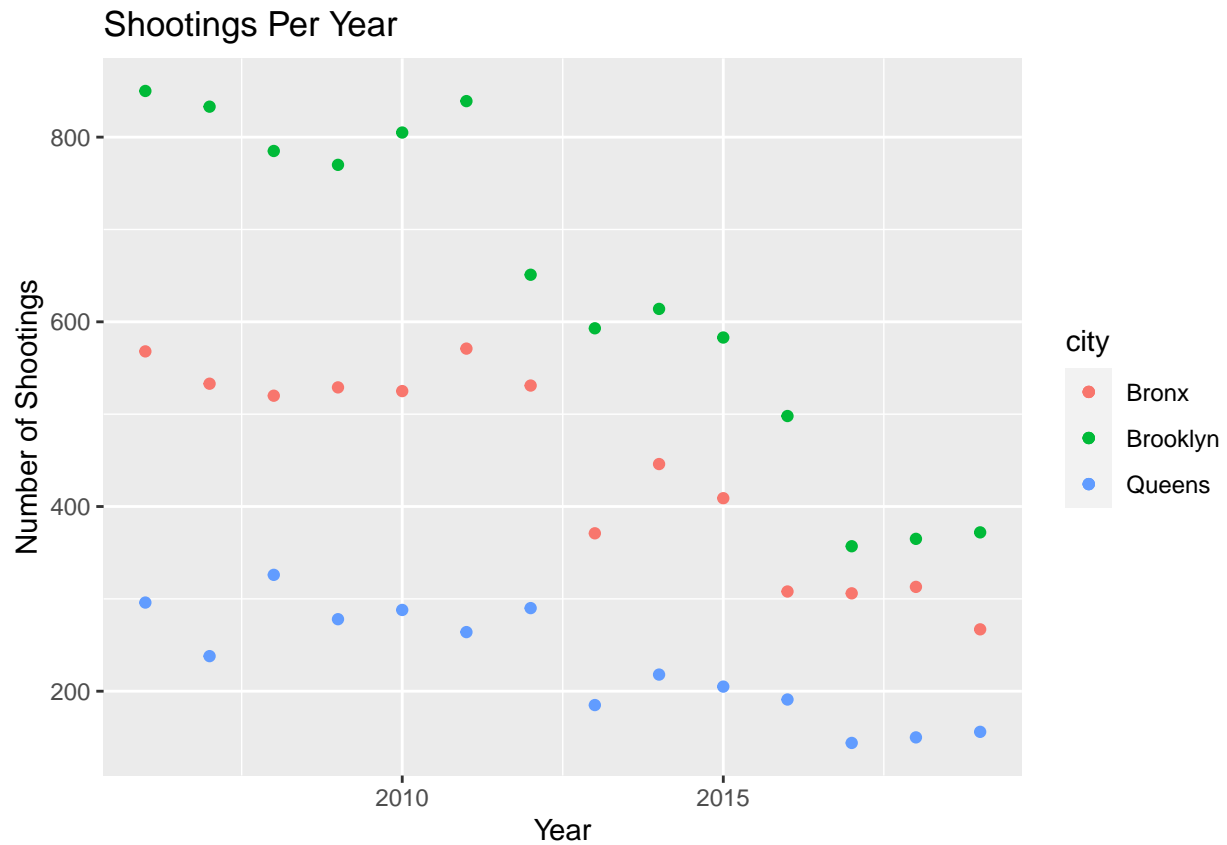
Data Modeling and Prediction

To get a better understanding of the overall trend of gun violence in these cities, we will use a prediction model to obtain a best fit line of the data. Reviewing this best fit line shows the rate that gun violence has been decreasing in these three cities. Another observation that can be seen from this data is that the rate of gun violence is in fact converging, which indicates a shift towards a potential rank shift.

The best method for confirming that a rank shift will occur in the future is to utilize our predction model and run it across a dataset which includes the years for the next decade.

```
ggplot(data_brook, aes(x = Year, y = freq, color = city)) +
  geom_point() +
  geom_point(data=data_queens, aes(color = city)) +
  geom_point(data=data_bronx, aes(color = city)) +
  geom_smooth(method='lm', formula=y~log(x)) +
  geom_smooth(data=data_queens, method='lm', formula=y~log(x)) +
  geom_smooth(data=data_bronx, method='lm', formula=y~log(x)) +
  labs(title = "Shootings Prediction", x = "Year", y = "Number of Shootings")
```
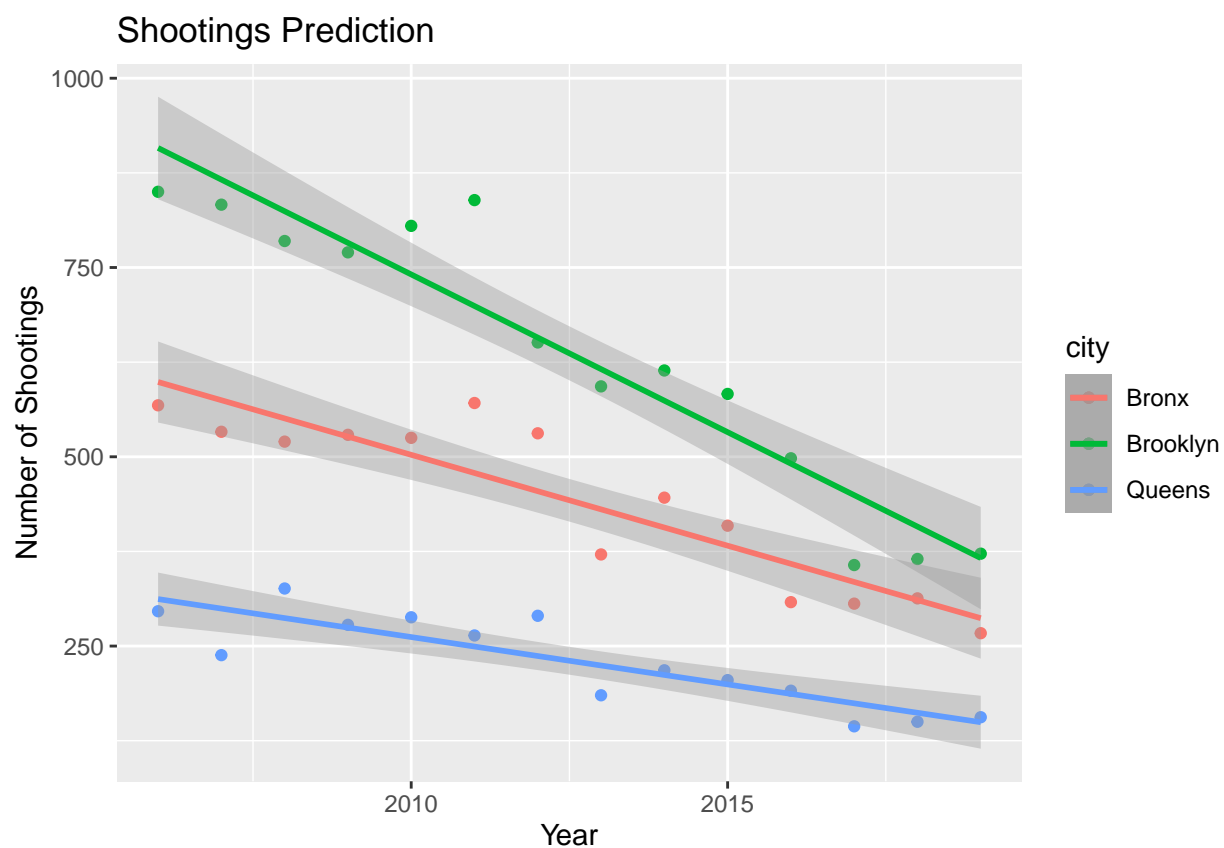
Future Analysis

What we see in this future prediction model is that the gun violence occurences continues to decrease over the next decade. However, there is some inaccuracy in this prediction model, due to the fact that gun violence goes negative by the year 2030. While this model may not provide an exact prediction, it is still useful in seeing how the rate of gun violence is expected to decrease. It also shows how the gun violence will change relative to eachother.

One key point we can see from this data is that gun violence in Brooklyn, which is currently the worst of the 3 cities, is expected to eventually improve to the point of becoming the least gun violent city of the group. Within the same 5 year span, its also predicted that Queens will reach the highest level of gun violoence. This 5 year span will range between 2026 and 2031 according to our prediction model.

```r
future_brook <- data.frame(Year=c(2020,2021,2022,2023,2024,2025,2026,2027,2028,2029,2030,2031,2032,2033)
lm_brook <- lm(freq~log(Year),data=data_brook)
future_brook <- future_brook %>% mutate(pred = predict(lm_brook,newdata = future_brook))
future_brook <- future_brook %>% mutate(city = "Brooklyn")

future_queens <- data.frame(Year=c(2020,2021,2022,2023,2024,2025,2026,2027,2028,2029,2030,2031,2032,203
lm_queens <- lm(freq~log(Year),data=data_queens)
future_queens <- future_queens %>% mutate(pred = predict(lm_queens,newdata = future_queens))
future_queens <- future_queens %>% mutate(city = "Queens")

future_bronx <- data.frame(Year=c(2020,2021,2022,2023,2024,2025,2026,2027,2028,2029,2030,2031,2032,2033)
lm_bronx <- lm(freq~log(Year),data=data_bronx)
future_bronx <- future_bronx %>% mutate(pred = predict(lm_bronx,newdata = future_bronx))
future_bronx <- future_bronx %>% mutate(city = "Bronx")

#Plot the future predictions
ggplot(future_brook, aes(x = Year, y = pred, color = city)) +
  geom_point() +
  geom_point(data=future_queens, aes(color = city)) +
  geom_point(data=future_bronx, aes(color = city)) +
  labs(title = "Shootings Future Predictions", x = "Year", y = "Number of Shootings")
```

Shootings Future Predictions
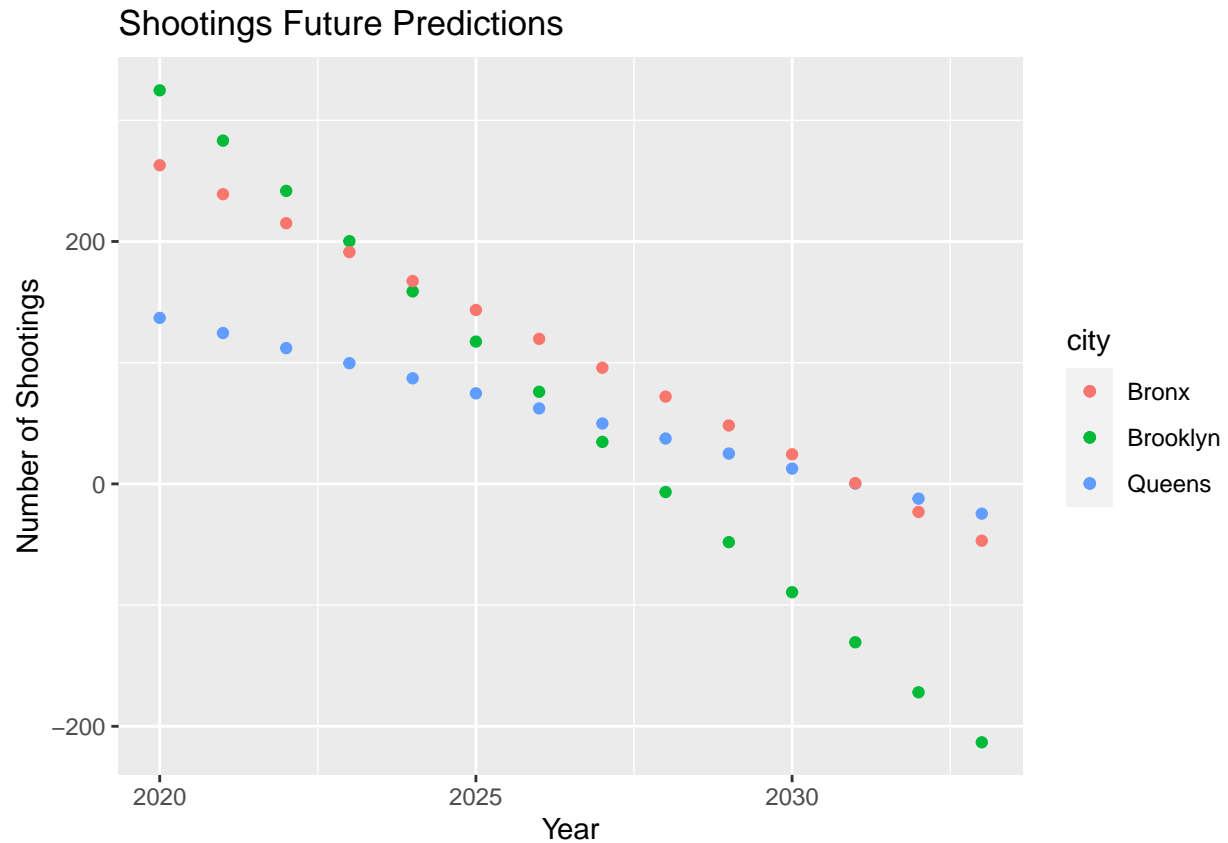
Sources of Error and Bias

There are a few main areas that may introduce error and bias into our analysis. The first is a source of bias, which was implemented to help remove outliers that were due in part to the current pandemic. During this analysis, we removed data for the year 2020, since shutdowns, business closures, loss of jobs, etc. may have led to an increase in mental illness and other factors that could lead to a significant increase in gun violence.

Another source of personal bias is related to the cities that I chose. There were many cities included in the data, but the cities that were chosen were familiar. Cities that were less familiar or didn't provide as many data points were left out of the analysis, even though they provide useful information in understanding how gun violence is in New York as a whole.

A source of error in the analysis is the prediction model used. We see in the future prediction that the number of occurences goes below 0, which we know is impossible. Based on this observation, we can determine that the prediction model is not an exact model. However, this model is still useful, and provides insight into how the different cities may change over time, and which cities may require more resources from the NYPD.

Summary

Throughout this analysis, we reviewed the data provided by the NYPD which shows the gun violence of different cities between the years of 2006 and 2020. We split the data for Brooklyn, Queens, and the Bronx. We then plotted the total shootings for each year and used a prediction model to determine the rate at which shootings was changing over the last decade, and also predict how shootings may change over the next decade.

This data provides some useful insight that could allow the NYPD to help understand where the efforts and resources should be targeted in the future to best meet the needs of cities and help fight gun violence.