DTSC 5301 NYPD Shooting Incidents Project

Karan Juneja, Andrew Edds, Darsh Shah, Vishwam Thakkar, & Rajas Umalkarm

September 21, 2021

Introduction

With the most guns per capita by several factors and increasing calls for gun policy reform, the United States' gun violence issue is significant and worth analyzing. Specifically in city centers, gun violence data has become readily available, allowing governmental organizations, police departments, researchers, and more to ask and answer relevant questions regarding gun violence dynamics, gun control, and crime reduction. This research project aims to explore these very dynamics in New York City, the largest US metropolitan. The research team is interested in gathering a variety of informative results concerning NYC's shootings, specifically with regard to how these crime patterns have differed during 2020.

2020 was a unique year, characterized by the worldwide novel COVID-19 pandemic, divisive elections in the United States, social unrest following ongoing police-sanctioned murdering of black people, and so much more. With such unexpected changes in social interactions and political dynamics, especially given the lockdowns that happened worldwide, there was an expectation that crime patterns would differ drastically from prior years. This project aims to uncover how NYC shooting incidents in 2020 were impacted by the unique set of issues that 2020 contained.

Whether it be the demographic backgrounds of perpetrators and victims or the spatial-temporal distribution of shootings across the five boroughs, there are several ways to explore how the unexpected social issues of 2020 affected shooting incidents in NYC. Using a clean and viable dataset, a set of informative visualizations, and statistical methods of modelling, this research project aims to answer the following question:

Did 2020 shooting patterns in New York City differ dramatically from prior years?

Data and Methods

The data for this analysis has been obtained from the New York Police Department and is titled "NYPD Shooting Incident Data (Historic)". The dataset outlines shooting incidents across the five boroughs from 2006-2020 with 23,568 observations, and each observation is a different shooting. The dataset includes many variables of interest, including perpetrator and victim demographics (age, race, sex, etc.), time of the shooting, latitude and longitude of the shooting, and more. As is common with many datasets, there are a number of biases embedded in the collection, presentation, and potential results of this analysis. These biases will be explored in the conclusion section.

To best understand how the set of unique challenges that 2020 contained impacted shooting incident patterns and dynamics, numerous visualizations and models have been crafted to analyze the dataset. Using R and RStudio software, several visualizations have been produced using the ggplot functions and tidyverse packages, and after identifying some dynamics of high interest, linear models have been created to understand how 2020 shooting incidents have differed from the prior years.

```
library(dplyr)
library(ggplot2)
library(anytime)
```

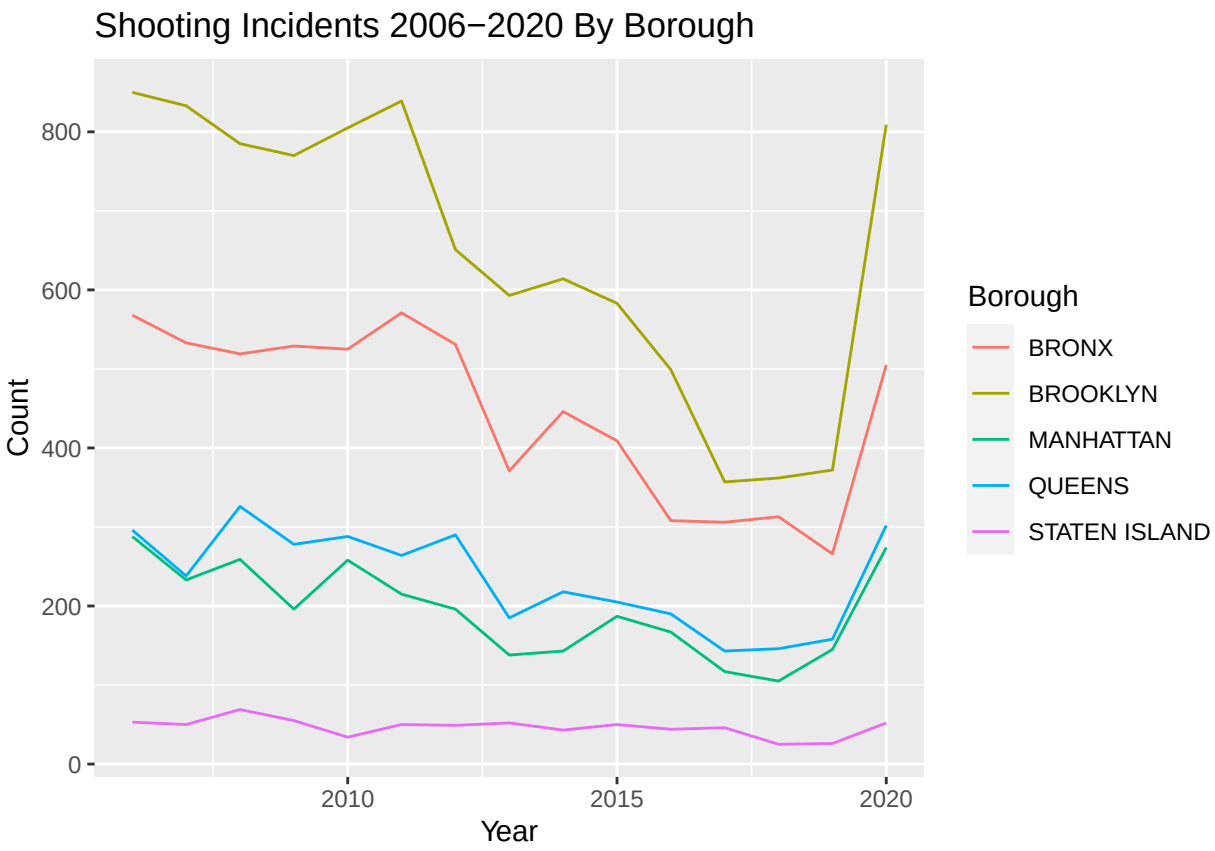
```
library(chron)
require(devtools)
library(tidyverse)
library(lubridate)
```

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

Visualizations

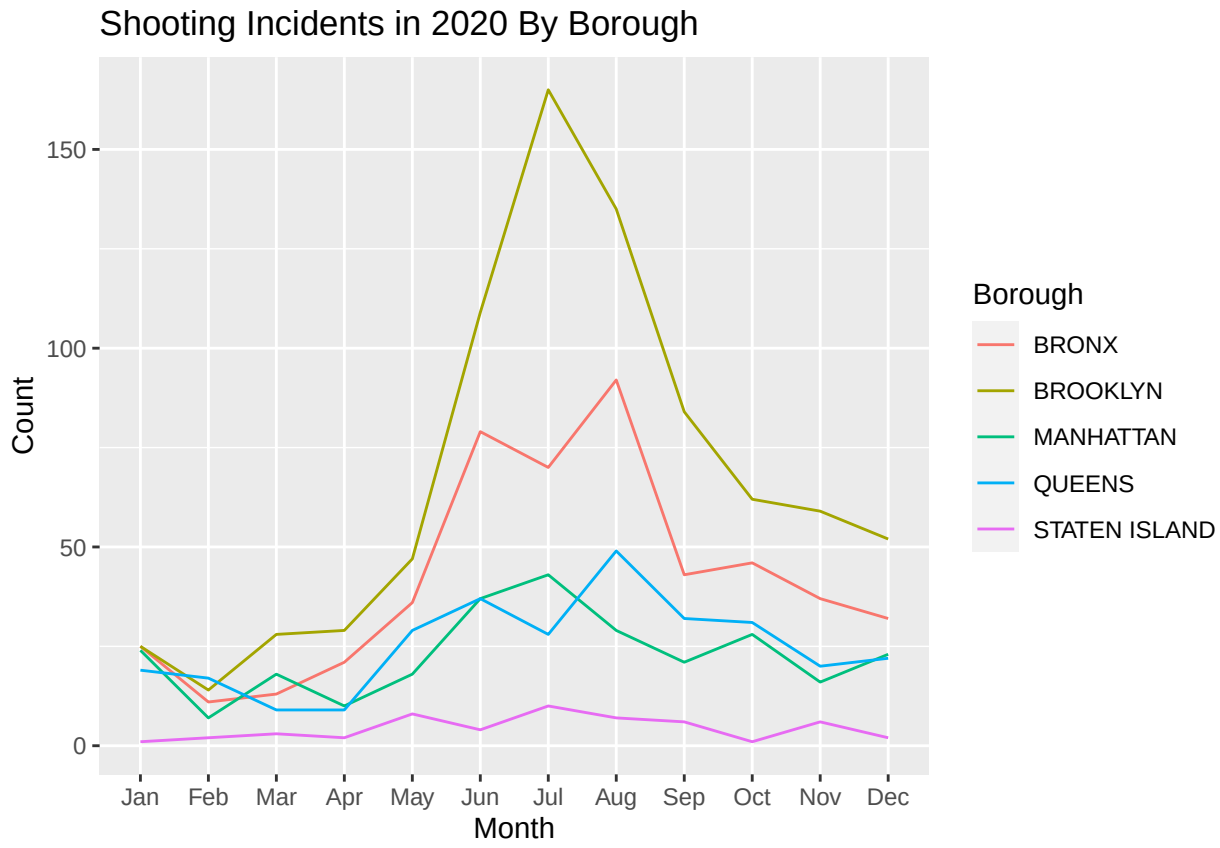
In conjunction with the time series analysis that would uncover differences in shooting patterns in 2020, several initial visualizations were created to understand the basic distribution, prevalence, and dynamics of NYC's shootings. From these visualizations, the research team had a greater understanding of the dataset, and thus could make informed decisions about which crime patterns could be interesting to explore with more visualizations and models. A couple of these initial visualizations are provided and explained below.

```
shootings_by_year <- shootings %>%
  group_by(BORO, year) %>%
  select(BORO, year) %>%
  mutate(count= n()) %>%
  distinct()

viz1 <- ggplot(shootings_by_year, aes(x=year, y=count, group=BORO, color=BORO)) + geom_line() + labs(ti
viz1
```



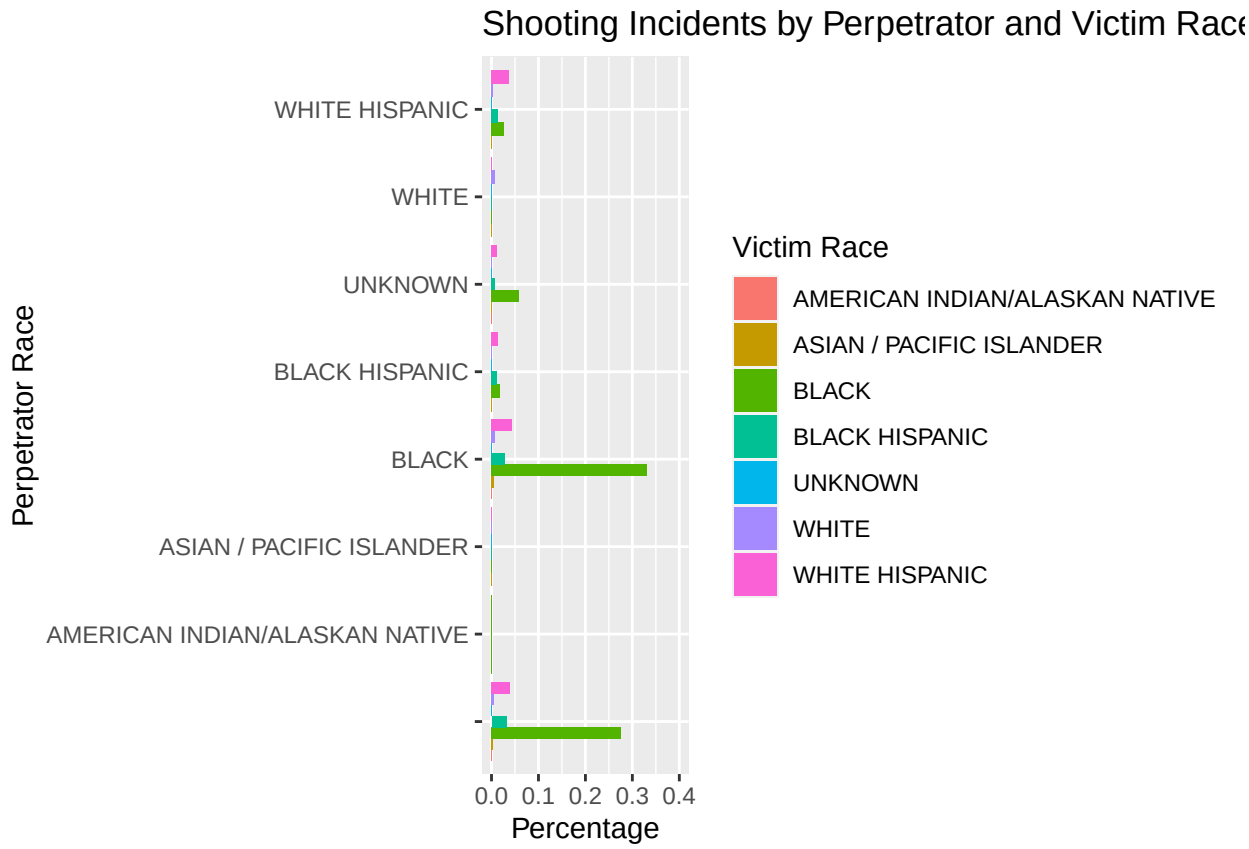
To get a general understanding of the temporal distribution of shootings across New York City, it was intuitive for the research team to begin by plotting all the shootings in the dataset grouped by borough. From this graph, the research team not only obtains an understanding of what areas of New York City have higher concentrations of shootings, but more importantly, an insight into the way shooting incidents spiked in 2020. By visualizing this spike, the research team obtained a sneak peek into how shooting incidents differed in 2020, intriguing the research team to dive deeper into this trend.

```
shootings_by_year_boro_2020 <- shootings %>%
  group_by(BORO, year, month) %>%
  select(BORO, year, month) %>%
  filter(year == 2020) %>%
  mutate(count= n()) %>%
  distinct()

viz2 <- ggplot(shootings_by_year_boro_2020, aes(x=month, y=count, group=BORO, color=BORO)) + geom_line()
viz2
```



After obtaining an understanding of how shooting incidents are spread across the entire dataset, the research team crafted this visualization to see the temporal distribution of shooting incidents in 2020 alone. From this graph, the research team can see that the rise of shooting incidents began around April, reached a high in every borough between July and August, and slowed down towards the end of the year. These insights correspond to several of the unique challenges contained in 2020, starting with the pandemic in April, the riots that carried through the summer, and the election that took place in November. From this visualization, the research team has a better understanding of how shooting incidents occurred during 2020, helping piece the puzzle together regarding why 2020 had such a large number of shootings in NYC.

```
viz3 <- shootings %>%
  count(PERP_RACE,VIC_RACE) %>%
  mutate(prop=n/sum(n)) %>%
  ggplot(aes(x=PERP_RACE,y=prop,fill=VIC_RACE))+ ylim(0,0.4)+
  geom_bar(stat = "identity",position = "dodge") + labs(title = "Shooting Incidents by Perpetrator and V
  ylab("Percentage")+ xlab("Perpetrator Race") + coord_flip()
viz3
```



In order to properly understand the nature of our dataset, it was necessary to craft numerous visualizations regarding the demographic background of both the perpetrators and victims of these shooting incidents. These plots included age, race, gender, and whether or not the victim died from the shooting. While each of these visualizations was interesting to examine, it was the race of perpetrator and victim visualization above that the research team wanted to include in this report. As is talked about in greater depth in the conclusion section, racism in policing is a topic that might be generating lots of bias in this dataset, and thus creating some display of this phenomenon was necessary for this analysis. While this visualization is not directly significant to the research team's goal of uncovering differences in shooting patterns between 2020 and prior years, the visualization provides an acknowledgment of this potential bias and a key insight into the construction of this dataset.

Model

After crafting numerous visualizations to picture the patterns and dynamics of shootings in NYC from 2006-2020, a model could be built to uncover whether or not 2020 had unexpected shooting incidents. Using the 2006-2019 period as a training dataset, the research team crafted a model to predict daily shootings in 2020

by aggregating shootings by month, and then compared this predictive model with the actual incidents of shootings in 2020. The model can be visualized below.

```
df_train <- df_final %>%
  filter(year <=  2019)
df_test <- df_final %>%
  filter(year == 2020)
```

```
lm_1 <- glm(Shootings~., data = df_train)
```

```
df_test <- df_test %>%
  mutate(predictions = predict(lm_1,df_test))
colors <-  c("Actual Shootings in 2020"='blue',"Shootings Before 2019"='green',"Predicted Shootings in 2
df_test %>% ggplot(aes(x = month , y = Shootings)) +
  geom_smooth(method='lm')+
  geom_smooth(aes(x = month , y = predictions, color= "predictions"),methond='lm')+
  geom_smooth(data=df_train,aes(x = month , y = Shootings),methond='lm',color='green')+
  labs(title = "Predictive Model of Shootings in 2020", x = "Month",
         y = "Shootings Per Day",
         color = "Legend")+
  scale_color_manual(values = colors)
```

The model above contains a set of graphs that display the daily aggregated shootings from 2006-2019 in green, the daily shootings of 2020 in blue, and the smoothed linear model prediction for daily shootings of 2020 in grey. This model has been used to show how the shooting incidents of 2020 differed from prior

years by comparing a prediction for 2020 based on prior years with the actual shooting incidents of 2020. From this model, we can see that the predicted shooting incidents in 2020 fall far below the actual shooting incidents of 2020, thus confirming the research groups hypothesis that shooting patterns in 2020 differed drastically from prior years.

Conclusion

From this set of visualizations and the predictive model for 2020 shooting incidents, it is clear that New York City shooting patterns in 2020 differed drastically from prior years. Given the unique set of challenges that the pandemic, election, and social unrest created during 2020, it is key to note that New York City saw higher levels of shooting incidents than have been experienced in nearly a decade. While these findings are helpful for understanding how health, justice, and political concerns impact crime dynamics in New York City, it is hard to know exactly how each of these scenarios contributed to the drastic rise in shooting incidents in 2020 without further analysis. Additionally, it is crucial that the potential biases contained within this dataset are acknowledged and understood.

Coming into this study, the research team acknowledged a personal bias regarding racism in policing in the United States. Police departments across the country have been criticized of racial discrepancies in the prosecution of people of color, leading to disparities in the allocation of resources and intentions of officers. There are many different ways that this bias has been constructed, but overall the research team had an expectation that the dataset would be biased in the racial demographics of perpetrators and victims of shootings in NYC. Given this bias, which the research group identified using visualizations, the research group focused on the time-series analysis of this data rather than the demographic data of potential perpetrators.

In addition to the personal bias acknowledged above, there were several aspects of the dataset's collection, construction, and presentation that must be recognized as potential dataset bias. Firstly, this dataset has been constructed to have each observation represent a different shooting incident, not a criminal, meaning that perpetrators can be recorded multiple times. Thus, analyses regarding the demographics of perpetrators and victims are skewed, and models that predict shooting incidents rather than criminal background. Another bias that might exist in the dataset is the lack of information regarding which shootings are recorded. The research group is curious about the potential recording of accidental shootings, shootings where the perpetrator is a police officer, and shootings where the perpetrator is lawfully self protecting or defending another person. It is difficult to prove the influence of these acknowledged potential biases, however, in future studies, results would be improved if these biases could be accounted for and properly understood.

Adjusting for these personal and dataset biases is not the only aspect of this analysis that could be improved with future studies. With larger datasets that could track COVID-19, social unrest, and political shifts, as well as less unknown values, better policy recommendations could be made and steps towards lowering violent crime could be taken. Nonetheless, this research project has uncovered several interesting dynamics in 2020 that altered the incidence of shootings in New York City, and the research group has gained a better understanding how global health crises, divisive political climates, and momentous social impact crime patterns in NYC.