UNIVERSITY OF MICHIGAN

ANALYSIS OF FATAL POLICE SHOOTINGS IN THE UNITED STATES

HEATHER JOHNSTON
KRITTIN TANGBORIBOONRAT
SHU ZHOU

May 19, 2022

1 Introduction

There has been growing concern about police brutality in recent years in the US, especially as it disproportionately affects black people and other people of color. The Black Lives Matter movement began in the wake of Michael Brown's death in 2014, and reached a new level of national recognition in May 2020 after George Floyd was murdered. A central element of this is when civilians are fatally shot by police officers in the line of duty. In this report, we study the pattern of fatal police shootings in the United States by analyzing the *Fatal Force* database, compiled by the Washington Post [Tate et al., 2021]. Other police incidents, such as shootings while the victim was in custody or civilians killed by police officers by means other than shooting are not available in this data. By using this data and additional data about US population, we perform an analysis of this dataset. We determine the demographic and geographical factors associated with police shootings and conduct exploratory analysis into the change in shooting rates since the murder of George Floyd. Additionally, we fit a prediction model based on the historical data.

2 Scientific questions

The scientific questions to be discussed are the following:

- **Monthly police shooting frequency per million people in each state** We combine the police shooting data with national population data to calculate the frequency of police shootings per million residents in each state. We examine the differences between the states and determine the states with lowest and highest police shooting rates.

- **Demographic characteristics among victims associated with police shootings.** A question of considerable importance is what demographic characteristics, such as race, predict becoming the victim of a police shooting. Since the data does not provide information about non-fatal police incidents, we instead examine which demographic traits predict that a victim of a police shooting will be unarmed and non-threatening. We apply random forest and logistic regression models to this data to understand which characteristics of a victim are associated with being unarmed.

- **Difference in average monthly police shootings before and after the George Floyd murder, separated by race of victim.** The murder of George Floyd was a tragedy that elevated the matter of police killings of black people to the cultural zeitgeist. Since polices and practices for police enforcement were changed in many places in the wake of George Floyd's death, we ask whether the average number of monthly police shootings has changed since May 2020, and whether this change differs by victim ethnicity group, gender and age groups.

- **Predicting the future number of police shootings from historical data.** We use a time-series analysis to predict future rates of police shooting.

3 Data and Methodology

The dataset *Fatal Force* is contains information about police shootings in the United States that have occurred since 2015. The dataset was built and is maintained by The Washington Post and is originally collected from police officers during the line of duty [Tate et al., 2021]. After acquiring the data, the Post also collects additional information from other sources such as local news reports and social media, which supply details about each shooting such as the race of the victims, the location of the shootings, and whether the police officers were wearing body cameras.

The dataset is frequently updated. It contains 16 variables and one unique identifier for each case. One challenge in the analysis of this dataset is that it contains only observations for lethal police shooting incidents and not any non-fatal police incidents, nor deaths by other means at the hands of police officers. This means that we cannot predict who will be more likely to be killed by police using this data alone. We need to relate this *Fatal Force* dataset with some other data or predict some subset of police shootings, such as which victims were unarmed. We also need to combine this data set with a national population data to calculate the average per capita number of shootings.

In order to examine the average number of police shootings per month per million people, we calculate the average number of monthly shootings (and associated 95% confidence intervals) for each state. Then we merge that with data with state populations from 2018 to calculate average monthly shootings per million residents per state [pop]. We report our findings in Section 4.1.

In order to determine which demographic traits are associated with a victim being disproportionately likely to be unarmed and non-threatening, we first define an indicator variable for the victims who were reported to be both listed as unarmed (or armed status not known) and not attacking at the time of death. We define five additional indicators for demographic traits and circumstances of the shooting, including whether the victim was male, a person of color, showed signs of mental illness, was not fleeing, and whether there is body camera footage. Using those five indicator variables and a continuous variable for the age of the victim at the time of the shooting, we fit a logistic regression model and a random forest model. The results are presented in section 4.2.

In order to determine whether the average number of monthly shootings has changed since George Floyd was killed, we create an indicator for whether the happened before or after the murder. For both before and after, we calculate the average number of monthly shootings for each racial category, and create 95% confidence intervals to evaluate significance. Our findings

are reported in section 4.3.

Finally, in order to fit a time series model, we use the 'ts' function in R. We use all data from 2015 through present to predict future rates of police shootings in the US. Our results are in section 4.4.

4 Results

4.1 Monthly police shooting per million population across states

In this section, we investigate the monthly police shootings per million residents across the different states. After finding the average number of police shootings every month in each state (and associated confidence intervals), we used the national population data in 2018 of all states to calculate the average number per million. Finally, we used a heat map to visualize the monthly police shootings per million population among states. The results are shown in the following figures.

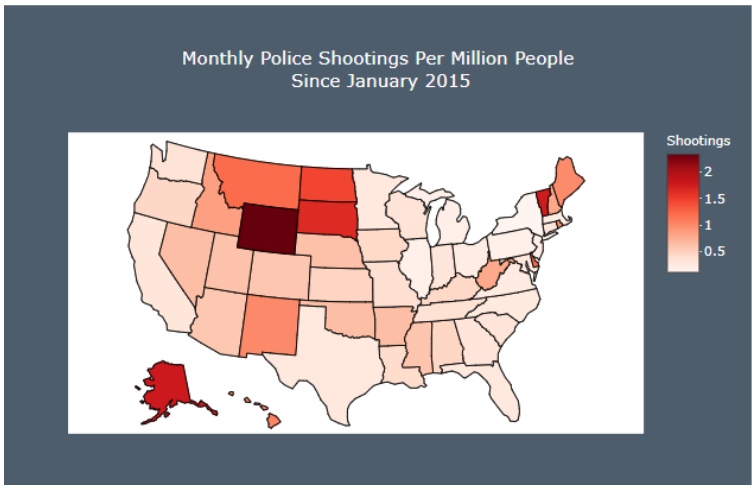

Figure 1: The Heat Map of the Monthly Police Shootings per 1 Million People among all states

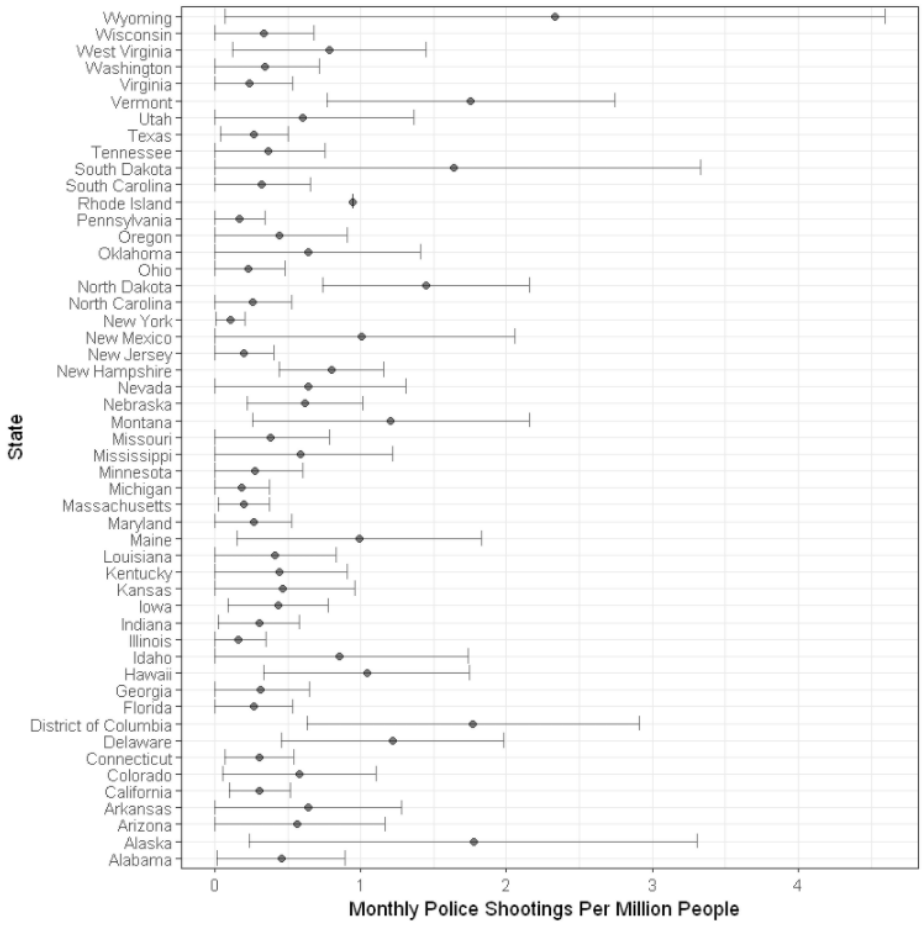

Figure 2: The 95% Confidence Intervals of the Monthly Police Shootings Per 1 Million People among all states

From the graphs, we can conclude that the states with the highest police shooting rates are Wyoming, Alaska, District of Columbia, Vermont, and South Dakota. The states with the lowest police shooting rates are New York, Illinois, Pennsylvania, Michigan, and New Jersey.

The result is somewhat counter-intuitive, since it is often reported that the police shootings frequently occur in major cities like Los Angeles, Chicago and New York. However, the states of New York, Illinois and California are among the states with the lowest police shooting rates. We hypothesize that the majority population of the suburban areas in those states is highly educated which would lead to a lower rate of crimes committed and would lower the probability of police shootings. On the other hand, states that are more rural with less population density, such as Wyoming and Alaska, the police shooting rates are much higher. This could be caused by the shootings on violations of traffic regulations since it is one of the the most common reasons for the police enforcement. In addition, for the state with more rural areas, there could be a higher chance

for violations of traffic regulations per a unit of population.

4.2 Unarmed Victims

In order to determine the predictors of someone being killed who does not pose an imminent threat to the police officers or others nearby, we fitted logistic regression and random forest models. For these models, we define five indicator variables to use, in addition to the continuous age variable.

The logistic regression model indicates that male victims are significantly less likely to be in the low-threat category, as well as victims showing signs of mental illness and senior people. Victims who are fleeing are significantly more likely to be in the low-threat category.

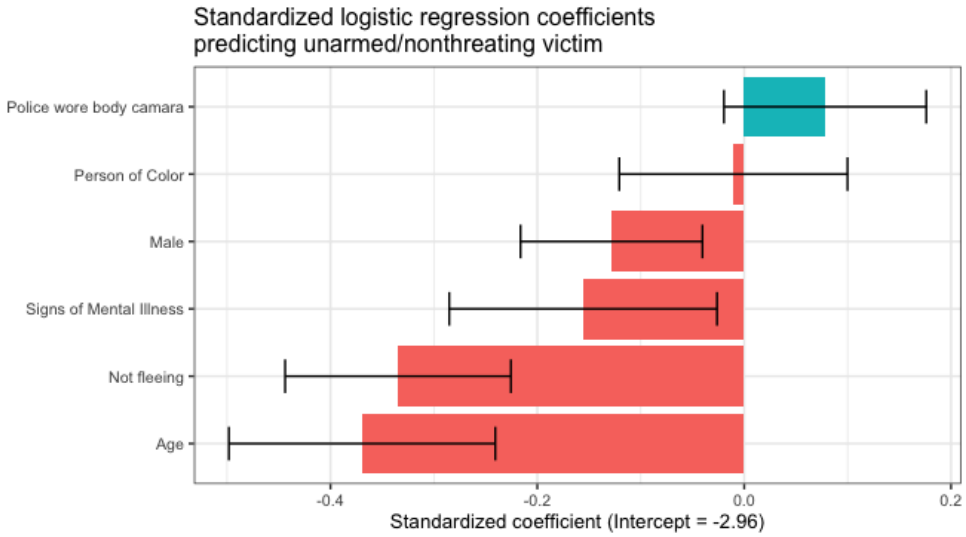

Figure 3: Standardized coefficient values (and 95% confidence intervals) for logistic regression model predicting whether a victim was unarmed.

The random forest variable importance as shown in Figure 4 indicates that age, gender, and signs of mental illness are the most important predictors.

Table 1: Logistic regression model predicting whether a victim will be in the low-threat category

	Dependent variable:
	not_attack
	Model
Age	−0.37 (0.07)***
Male	−0.13 (0.04)***
Person of Color	−0.01 (0.06)
Signs of mental illness	−0.16 (0.07)**
Not fleeing	−0.33 (0.06)***
Police wore body camera	0.08 (0.05)
Constant	−2.96 (0.06)***
Observations	6,467
Log Likelihood	−1,357.41
Akaike Inf. Crit.	2,728.83
Note:	*p<0.1; **p<0.05; ***p<0.01

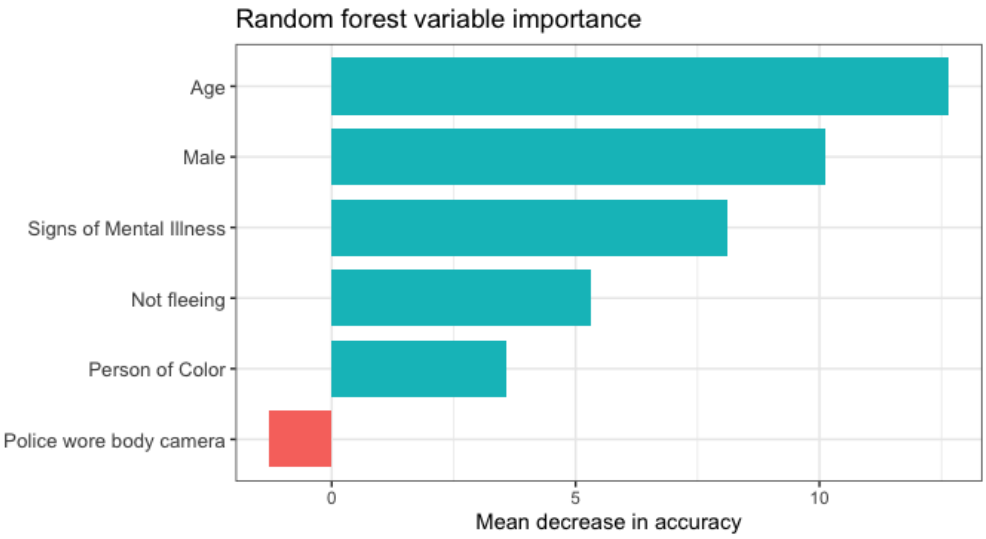

Figure 4: Random forest model mean decreases in the model accuracy for each predictor

4.3 The effect of the George Floyd Murder to the police shootings

We first divided our data set into two parts, the first part is for the cases before the George Floyd Murder, and the second part is for the cases after the George Floyd Murder. And we calcu-

lated the average number of murders among different races and also constructed the 95% confidence intervals from the means and standard errors obtained.

The result graph is shown below.

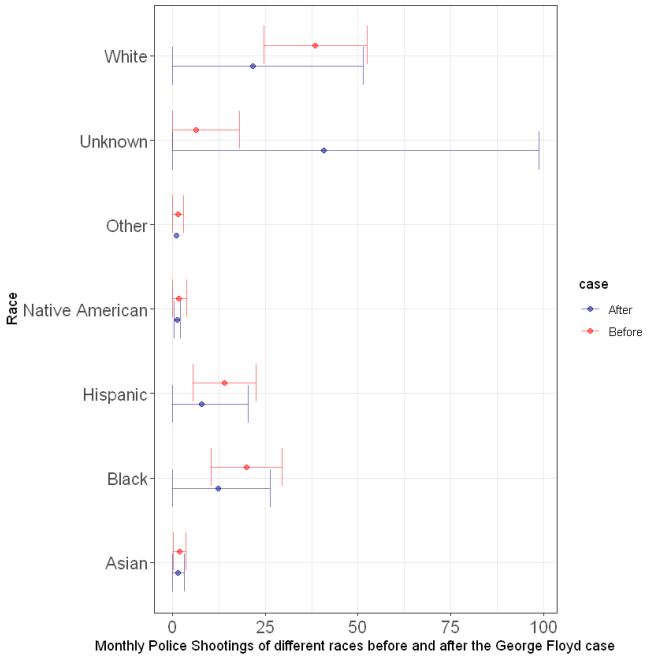

Figure 5: The 95% Confidence intervals of the Monthly Police Shootings among different ethnicity groups before and after the George Floyd Murder

From this graph, it is clear that the number of police shootings have decreased among all ethnicity groups after the George Floyd Murder (except for the Unknown ethnicity groups), and the decrease in all the other ethnicity groups are significant. And for the unknown ethnicity groups, the high variance is caused by lacking in data points.

Then we construct the similar confidence interval regarding the gender and age groups.

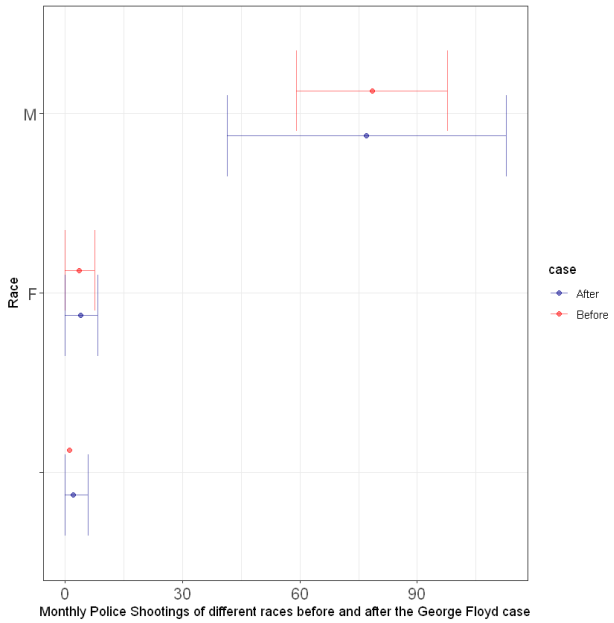

Figure 6: The 95% Confidence intervals of the Monthly Police Shootings among different genders before and after the George Floyd Murder

From the graph above, we cannot verify any significant differences between the monthly shootings before and after the George Floyd Murder with regard to gender.

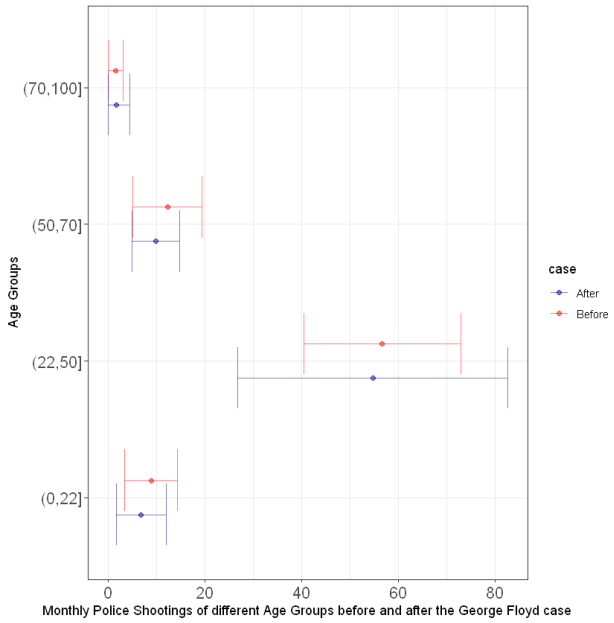

Figure 7: The 95% Confidence intervals of the Monthly Police Shootings among different age groups before and after the George Floyd Murder

From the graph above, a decrease can be indicated from the age groups for almost all the age groups except for the group of (70,100]. And the decrease is significant for the (0,22] and (50,70] age groups. Though the decrease is not statistically significant for the group of (22,55], we can still use this as a reference of decreasing.

4.4 Time Series Analysis

After investigating into the cases of George Floyd, it would be an interesting topic to verify the changes of police shootings over months after the murder has taken place.

We first converted the data into the time series. Before looking at the trend and changes, we should check the seasonality .

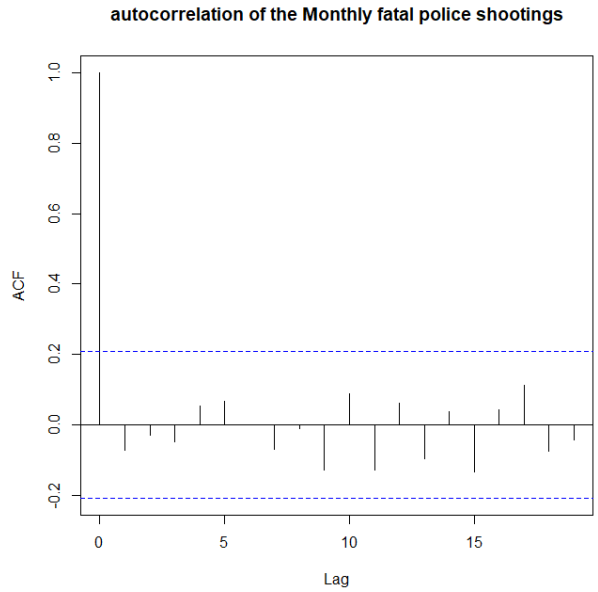

Figure 8: The autocorrelation plot of monthly fatal police shootings in the USA

Then we look at the polarized plot the monthly fatal police shootings changes from January 2019 to May 2022.

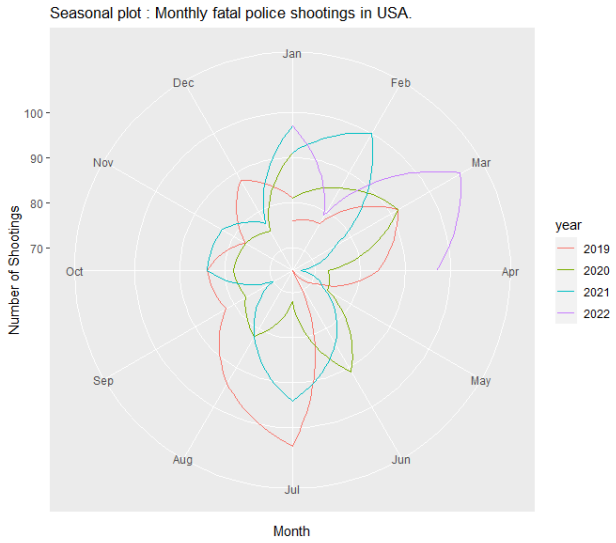

Figure 9: Monthly fatal police shootings in the USA

According to area covered by the plot above, it could be verified that the number of police shootings in the United States does decrease in 2020 and 2021. However the number of police shootings then increased again from the beginning of 2022.

To further dig into this analysis of the impact of George Floyd murder, we should dig into the case of Black ethnicity:

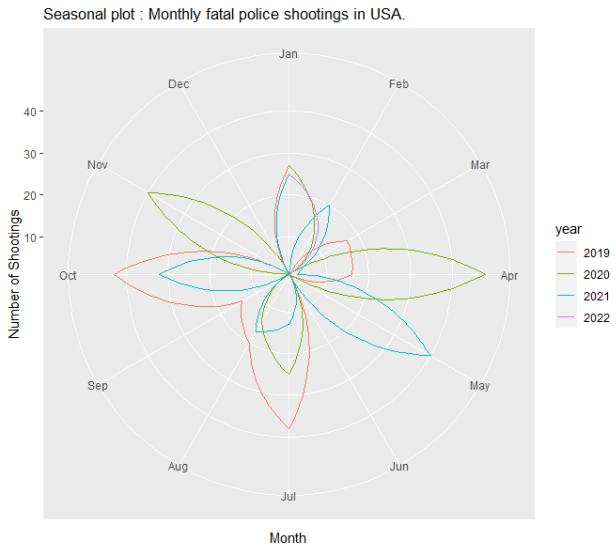

Figure 10: Monthly fatal police shootings of black ethnicity in the USA

For black ethnicity group, a immediate decrease can be noticed after May 2020. Also, according to the area covered by the line of each year, we can verify a steady trend of decreasing in the number of shootings. Also, the monthly shootings number has not exceeded 30 since the murder had taken place. This result indicates a good impact of the Black Live Matter movement to the criterion of the law enforcement of the police.

Them we calculated the prediction intervals with the levels of 0.80 and 0.95 for the next several years based on the current data for each month. The prediction intervals is constructed through the 'stlf' function of r.

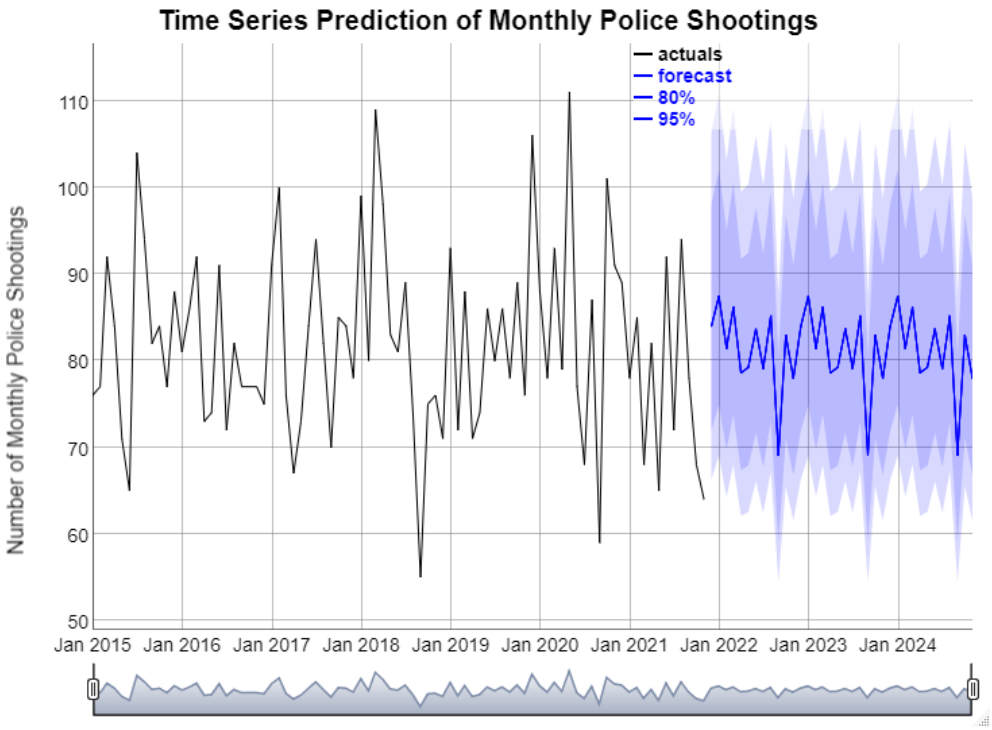

Figure 11: The 80% and 95% Prediction Intervals of the Monthly Police Shootings for the next several years

We can see that the number of the shootings still varies a lot among months, but the general trend of shooting numbers would indicate a decreasing trend for the next few years based on the prediction intervals.

5 Conclusion

Our results does provide sufficient evidence to conclude that the police shooting rate has decreased significantly since George Floyd was murdered. And it is possible that police precincts have changed their polices or practices and that the rate of fatal police shootings will decline in the next several years, especially when treating the Black ethnicity group. However, the time series analysis shows too large a confidence interval to predict how police shooting rates will change in the future. Moreover, we also have learned about the predictors and factors that led to the shootings among the unarmed victims using the logistic regression and random forest models. In the future, if more data is to be collected, we would like to explore more details in the other causes of victims' deaths such as shootings in custody or other means rather than shooting. We would like to explore more statistical methods that could provide reasonable predictions of the aforementioned topics in further research.

References

Julie Tate, Jennifer Jenkins, Steven Rich, and John Muyskens. Fatal Force. The Washington Post, 2021. URL https://github.com/washingtonpost/data-police-shootings.

US State populations. URL https://www.kaggle.com/lucasvictor/us-state-populations-2018.

6 Appendix: Code

```
library(forecast)
library(dygraphs)
library(tidyverse)
library(randomForest)
library(stargazer)
library(plotly)
################################################################################
# Data cleaning and prep
################################################################################
police_shootings <- read.csv(file = 'fatal-police-shootings-data.csv')
police_shootings$case = 1
police_shootings$date <- as.Date(police_shootings$date)

police_shootings$months <- format(police_shootings$date, "%Y-%m")
monthly_shootings_state <- police_shootings %>%
  group_by(months, state) %>%
  summarize(monthly_shootings = sum(case))

#create monthly Shooting point estimates and standard deviations
monthly_shootings_pe <- monthly_shootings_state %>% group_by(state) %>%
  summarize(avg_shootings = mean(monthly_shootings),
            sd_shootings = sd(monthly_shootings))%>%
  rename(code = state)%>%
  merge(state_populations, by = "code") %>%
  mutate(monthly_shootings_million = avg_shootings/pop_2014 * 1000000) %>%
  mutate(sd_shootings_million = sd_shootings/pop_2014 * 1000000 )

monthly_shootings_pe <- monthly_shootings_pe[,c("code",
```

```
  "monthly_shootings_million", "sd_shootings_million","state" )]

monthly_shootings_pe[order(
  monthly_shootings_pe$monthly_shootings_million),]

################################################################################
# Averages by state
################################################################################
# specify some map projection/options
g <- list(scope = 'usa', projection = list(type = 'albers usa'))
titlefont<- list(color = "white")

# Create a map of the police shootings per million people by state
plot_geo(monthly_shootings_pe, locationmode = 'USA-states') %>%
  add_trace(z = ~monthly_shootings_million, text = text,
  locations = ~code,
            color = ~monthly_shootings_million,
            colors = 'Reds', hoverinfo = 'text') %>%
  colorbar(title = "Shootings", titlefont = titlefont,
           tickfont = titlefont, tickcolor = "white") %>%
  layout(title = 'Monthly Police Shootings Per Million People\n
    Since January 2015', geo = g, autosize = F,
         paper_bgcolor = "#4e5d6c", titlefont = titlefont,
         margin = list(t = "110"))

## Create 95% Confidence Intervals
alpha = qnorm(.975)
monthly_shootings_ci = monthly_shootings_pe %>%
  mutate(lwr = pmax(monthly_shootings_million - alpha * sd_shootings_million,0),
  upr = monthly_shootings_million + alpha * sd_shootings_million)

monthly_shootings_ci

monthly_shootings_ci %>% ungroup() %>%
  ggplot( aes(y = monthly_shootings_million, x = state)) +
  geom_point( position = position_dodge(width = .5), alpha = 0.5 ) +
```

```
  geom_errorbar( aes(ymin = lwr, ymax = upr), alpha = 0.5,
                 position = position_dodge(width = .5)
  ) +
  coord_flip() +
  theme_bw() +
  scale_color_manual( values =  "red") +
  ylab('Monthly Police Shootings Per Million People  ' ) +
  xlab('State')

################################################################################
# George Floyd
################################################################################

police_shootings$date <- as.Date(police_shootings$date)

after_pandemic <- police_shootings[(police_shootings$date> "2020-5-25"),]
after_pandemic$pandemic = "After"
before_pandemic <- police_shootings[(police_shootings$date<= "2020-5-25"),]
before_pandemic$pandemic = "Before"

police_shootings_p <- rbind(after_pandemic, before_pandemic)

monthly_shootings_race <- police_shootings_p %>%
  group_by(months,pandemic, race) %>%
  summarize(monthly_shootings = sum(case))

monthly_shootings_race_pe <- monthly_shootings_race %>%
  group_by(pandemic,race) %>%
  summarize(avg_shootings = mean(monthly_shootings),
    sd_shootings = sd(monthly_shootings))

monthly_shootings_race_pe$race <- sub("^$", "Unknown",     monthly_shootings_race_pe$race)
monthly_shootings_race_pe[c("race")][
  which(monthly_shootings_race_pe$race == "A"),] <- "Asian"
monthly_shootings_race_pe[c("race")][
  which(monthly_shootings_race_pe$race == "B"),] <- "Black"
```

```
monthly_shootings_race_pe[c("race")][
  which(monthly_shootings_race_pe$race == "H"),] <- "Hispanic"
monthly_shootings_race_pe[c("race")][
  which(monthly_shootings_race_pe$race == "N"),] <- "Native American"
monthly_shootings_race_pe[c("race")][
  which(monthly_shootings_race_pe$race == "W"),] <- "White"
monthly_shootings_race_pe[c("race")][
  which(monthly_shootings_race_pe$race == "O"),] <- "Other"

alpha = qnorm(.975)
monthly_shootings_race_ci = monthly_shootings_race_pe %>%
  mutate(lwr = pmax(avg_shootings - alpha * sd_shootings,0),
  upr = avg_shootings + alpha * sd_shootings)

monthly_shootings_race_ci$case <- as.factor(monthly_shootings_race_ci$pandemic)

monthly_shootings_race_ci %>%
  ggplot( aes(y = avg_shootings, x = race, color = `case`)) +
  geom_point( position = position_dodge(width = .5), alpha = 0.5 ) +
  geom_errorbar( aes(ymin = lwr, ymax = upr), alpha = 0.5,
                 position = position_dodge(width = .5)
  ) +
  coord_flip() +
  theme_bw() +
  theme(axis.text.x=element_text(size=rel(1.5))) +
  theme(axis.text.y=element_text(size=rel(1.5))) +
  scale_color_manual( values = c("darkblue", "red")) +
  ylab('Monthly Police Shootings of different races before and
    after the George Floyd case' ) +
  xlab('Race')

##############################################################################
# Predicting unarmed and nonthreatening status
##############################################################################
url1 <- "https://raw.githubusercontent.com/washingtonpost/data-police-shootings"
url2 <- "/master/fatal-police-shootings-data.csv"
```

```
police <- read.csv(paste0(url1, url2), stringsAsFactors = TRUE)

police$date <- as.Date(police$date)
police$race <- relevel(police$race, ref = "W")
police$not_white <- police$race != "W"
police$male <- police$gender == "M"
police$not_flee <- police$flee == "Not fleeing"
police$low_threat <- police$threat_level != "attack" &
  (police$armed == "unarmed" | police$armed == "")

police_std <- data.frame(age = scale(police$age),
                         male = scale(as.numeric(police$male)),
                         not_white = scale(as.numeric(police$not_white)),
                         signs_of_mental_illness = scale(as.numeric(
                           police$signs_of_mental_illness)),
                         not_flee = scale(as.numeric(police$not_flee)),
                         body_camera = scale(as.numeric(police$body_camera)),
                         not_attack = police$not_attack)

model <- glm(not_attack ~ age + male +
  not_white + signs_of_mental_illness + not_flee + body_camera,
    data = police_std, family = "binomial")

model_info <- data.frame(values = model$coefficients,
                         se = summary(model)$coefficients[ , 2],
                         names = names(model$coefficients))
model_info$long_names = c("Intercept","Age",
                          "Male",  "Person of Color",
                          "Signs of Mental Illness","Not fleeing",
                          "Police wore body camara")
model_info$lwr <- model_info$values - 1.96*model_info$se
model_info$upr <- model_info$values + 1.96*model_info$se
model_info <- model_info %>% filter(long_names != "Intercept")

ggplot(model_info, aes(x = reorder(long_names, values), y = values,
  fill = as.factor(sign(values)))) +
```

```
  geom_bar(stat = "identity") +
  geom_errorbar(mapping = aes(ymin = lwr, ymax = upr), width = .5) +
  labs(title = "Standardized logistic regression coefficients\n
    predicting unarmed/nonthreating victim",
       x = "",
       y = "Standardized coefficient (Intercept = -2.96)") +
  theme_bw() +
  theme(legend.position = "none", text = element_text(size = 10)) +
  coord_flip()

stargazer(model,
          title="Logistic regression model predicting whether a
          victim will be in the low-threat category",
          type = "latex",
          float = TRUE,
          report = "vcs*",
          #  se=lapply(models, function(x) sqrt(diag(vcovHC(x, type = "HC1")))),
          no.space = TRUE,
          header=FALSE,
          single.row = TRUE,
          font.size = "small",
          intercept.bottom = TRUE,
          covariate.labels = c("Age", "Male", "Person of Color",
                               "Signs of mental illness", "Not fleeing",
                               "Police wore body camera"),
          column.labels = c("Model"),
          column.separate = c(1,1,1,1),
          digits = 2,
          t.auto = F,
          p.auto = F,
          notes.align = "l",
          notes.append = TRUE,
          df = FALSE
)

police_rf <- police %>%
```

```
  select(low_threat, age, male, not_white, signs_of_mental_illness, not_flee,
         body_camera) %>%
  drop_na()
rf_model <- randomForest(
  as.factor(low_threat) ~ age + male + not_white +
  signs_of_mental_illness + not_flee + body_camera,
  data = police_rf,
  importance = TRUE
)

var_importance = varImpPlot(rf_model)

var_importance %>% as.data.frame() %>%
  mutate(vars = c("Age", "Male", "Person of Color", "Signs of Mental Illness",
                  "Not fleeing", "Police wore body camera")) %>%
  ggplot( aes(x = reorder(vars, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy,
              fill = as.factor(sign(MeanDecreaseAccuracy)))) +
  geom_bar(stat = "identity", width = .8) +
  labs(title = "Random forest variable importance",
       y = "Mean decrease in accuracy",
       x = "") +
  scale_fill_discrete(c("darkblue", "red")) +
  theme_bw() +
  theme(axis.text.y = element_text(size=10),
        axis.title.y = element_blank(),
        legend.position = "none") +
  coord_flip()

##############################################################################
# Time series model
##############################################################################
monthly_shootings_state <- police_shootings %>%
  group_by(months) %>%
  summarize(monthly_shootings = sum(case))

monthly_shootings_state_ts<-ts(monthly_shootings_state$monthly_shootings,
```

```
  start=c(2015,1),
  end=c(2021,11),
  frequency=12)

monthly_shootings_state_ts %>%
  stlf(lambda = 0, h = 36) %>%
  {cbind(actuals=.$x, forecast_mean=.$mean,
         lower_95=.$lower[,"95%"], upper_95=.$upper[,"95%"],
         lower_80=.$lower[,"80%"], upper_80=.$upper[,"80%"])} %>%
  dygraph(main="Time Series Prediction of Monthly Police Shootings",
  ylab = "Number of Monthly Police Shootings") %>%
  dyAxis("y", valueFormatter = interval_value_formatter) %>%
  dySeries("actuals", color = "black") %>%
  dySeries("forecast_mean", color = "blue", label = "forecast") %>%
  dySeries(c("lower_80", "forecast_mean", "upper_80"),
           label = "80%", color = "blue") %>%
  dySeries(c("lower_95", "forecast_mean", "upper_95"),
           label = "95%", color = "blue") %>%
  dyLegend(labelsSeparateLines=TRUE) %>%
  dyRangeSelector() %>%
  dyOptions(digitsAfterDecimal = 1) %>%
  dyCSS(textConnection(".dygraph-legend {background-color:
  rgba(255, 255, 255, 0.5) !important; }"))
```