Statistical Data Analysis on fatal police shootings in the US

Final Project

Hiu Man (Stephanie) Lam
ID: 913358632

Dhanya Chandrasekhar
ID: 999177426

Janet Loyola
ID: 912092672

Juliana Noronha
ID: 998754002

Abstract
The purpose of our project is to analyze the Washington Post Fatal Police Shooting data set and determine any correlations between the given variables. In particular, we'd like to see how gender, race, threat level and whether the victim is armed relate to the likelihood of being fatally shot by police.

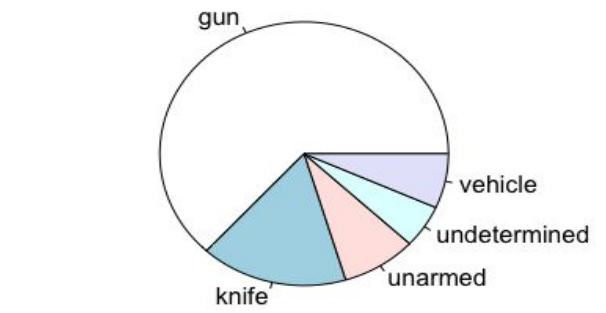
I.Introduction

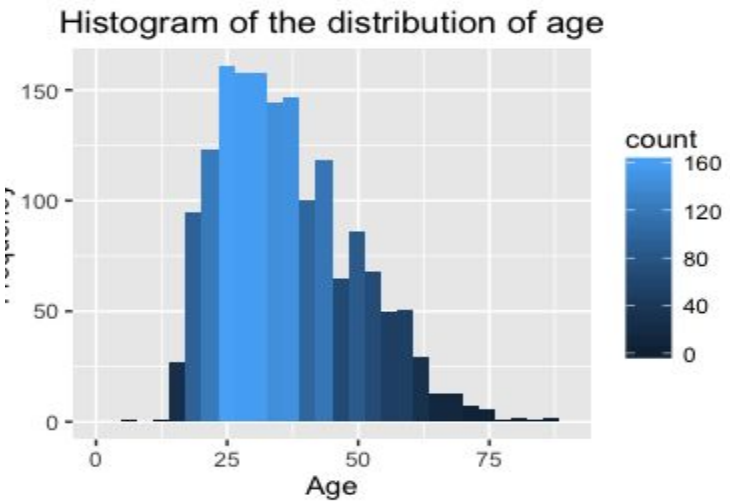
The project data is coming from a GitHub database, compiled by the Washington Post, on all shootings in which a police officer, in the line of duty, shot and killed a civilian in the United States in 2015 and 2016. Variables such as race, gender, signs of mental illness are included in the data set. This database came about as form of tracking police activity and to increase accountability among police departments nationwide. Our data set contains a total of 14 variables including name, date, manner of death, armed, age, gender, race, city, state, signs of mental illness, threat level, flee, body camera.

II.Interpretation

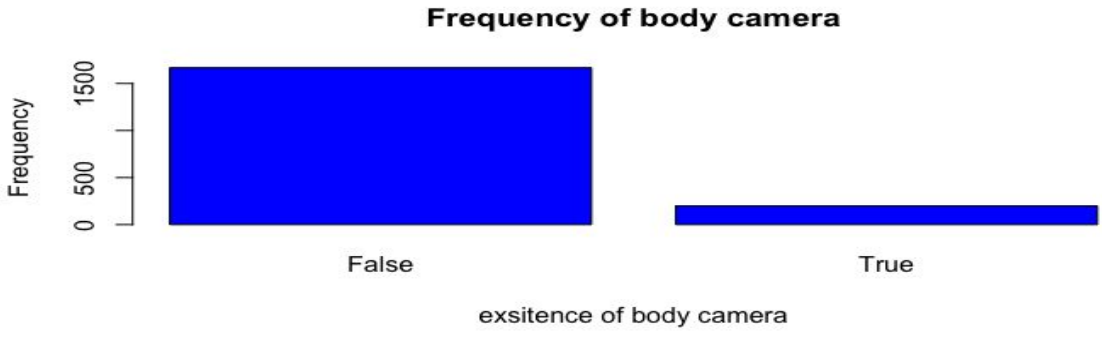
Body Camera and Armed Analysis

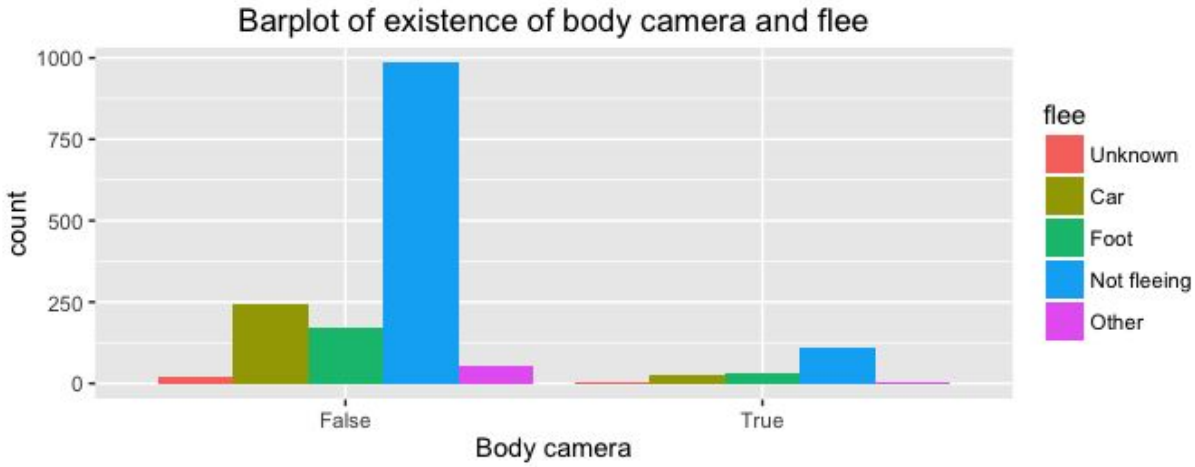
Since there are 59 different variables in the category "armed", which indicates what weapon was held by the subject at the time of the shooting, it is very hard to extract useful information. Therefore, we will focus on the top five armed categories, including gun, knife, unarmed, vehicle, and undetermined. By conducting a statistical summary, analysis shows that less than 10 percent of fatal shooting by police involved an unarmed civilian.



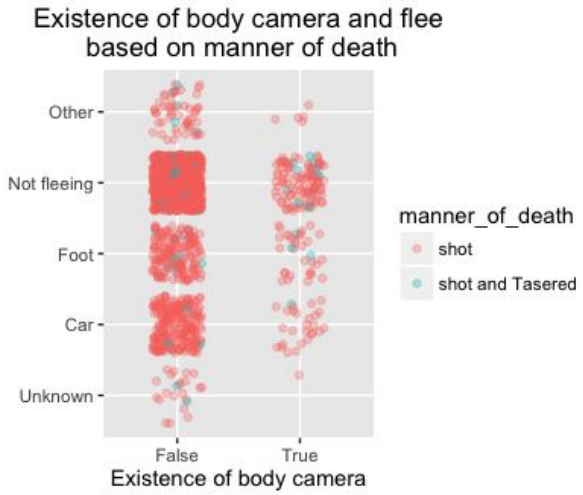


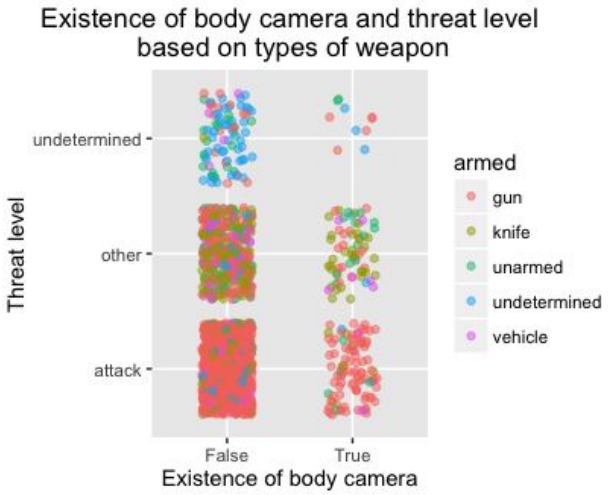
Since we are interested in whether the existence of a body camera would affect shooting, we used various plots to help examine the relationship between body camera and other categorical variables such as manner of death, flee, and threat level etc. Based on the histogram and barplot below, we can see that if the police officer was not using a body camera, there was a significantly higher shooting frequency. Significantly, police officers shot the most often if they were not wearing a body camera and the suspect was not fleeing.

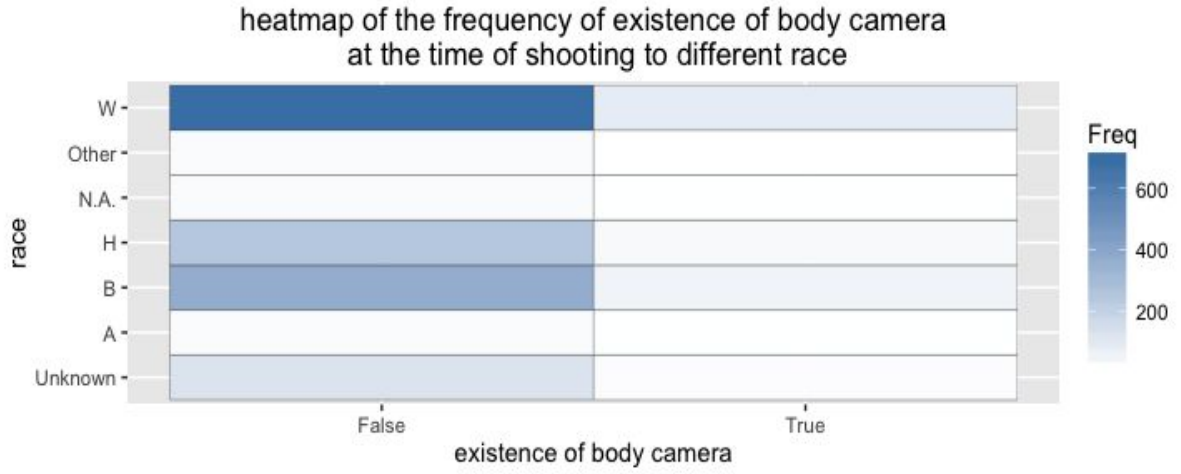




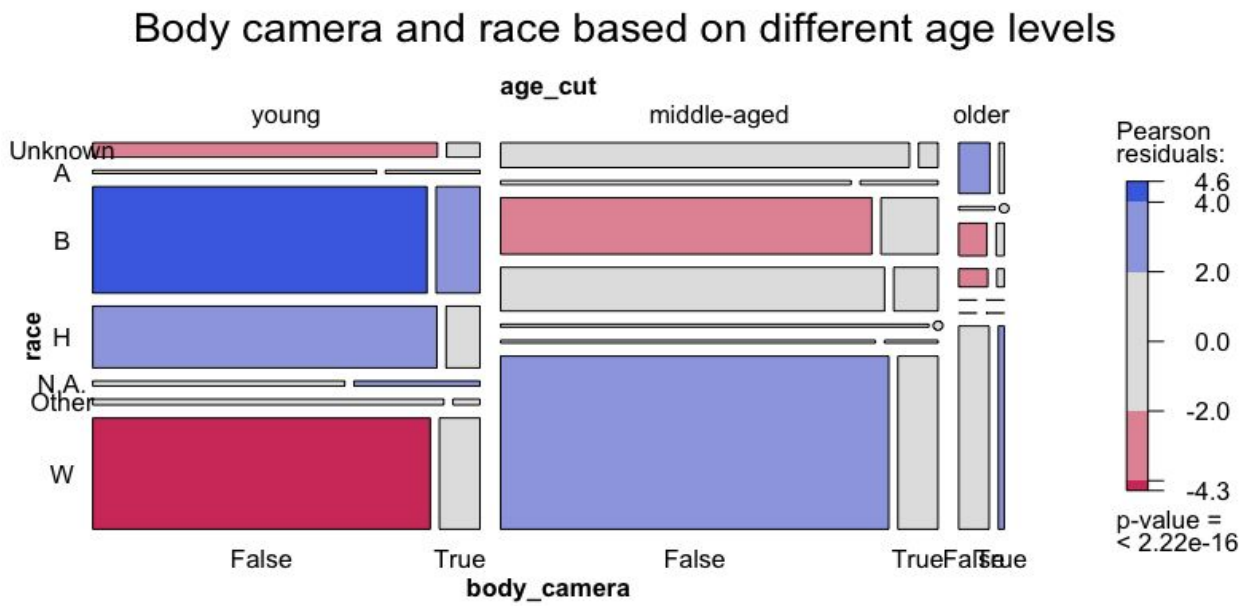
From the two scatter plots, it is also very clear that police would fire much more often when they were not wearing a body camera, especially when the person is holding a gun, which is also when the threat level is considered as attack.





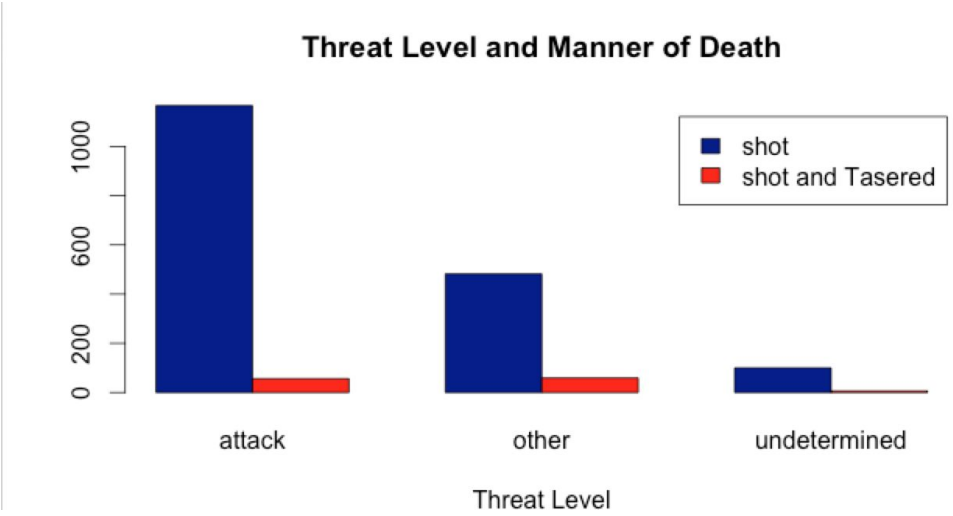


The heatmap above shows that white people are shot much more frequently compared to other races. Therefore, we would like to further examine the existence of the body camera effect on race. We divided age into 3 groups, young(6-33), middle-aged(33-60), and older (60-86) and used the mosaic plot below to visualize the relationships across the age groups:



The result actually implies that younger black people have a much higher rate of getting shot, regardless of body camera usage. On the other hand, young hispanic people are much more when police was not wearing a camera, while young native americans have a higher rate when there was a camera. For middle-aged class, there are more observations than expected for the whites to be shot when the policeman was not wearing a body camera. And police with body camera seems to fire more easily to the older whites too. Generally, asians are less likely to involve in the fatal shooting.

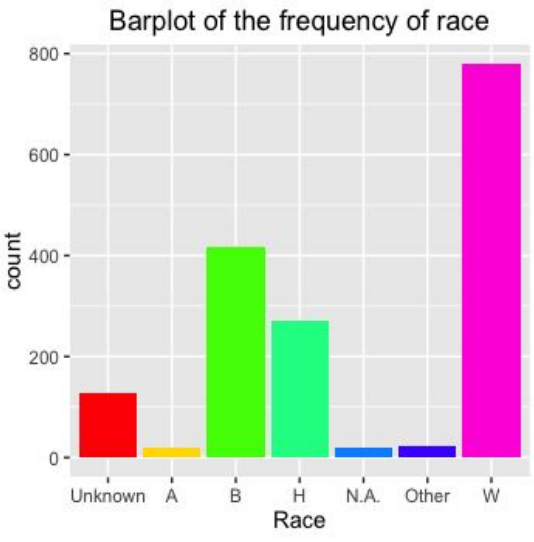
Threat Level and Manner of Death Analysis

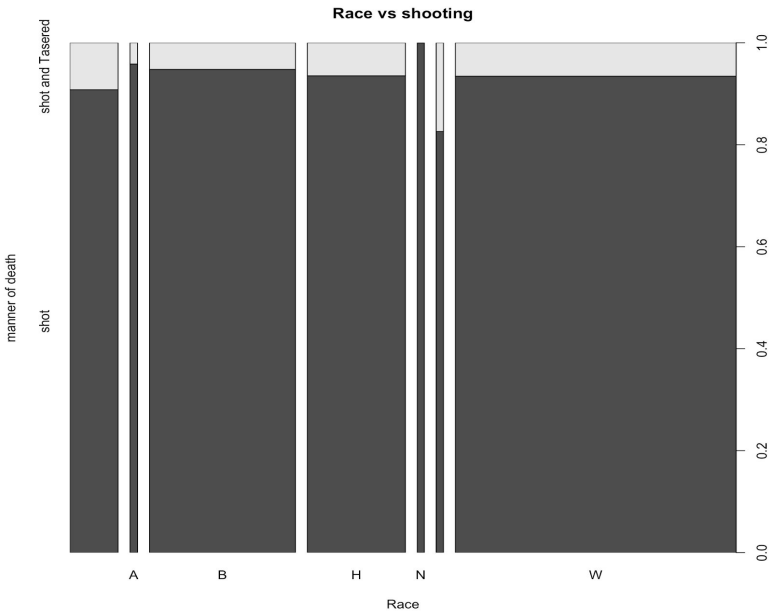


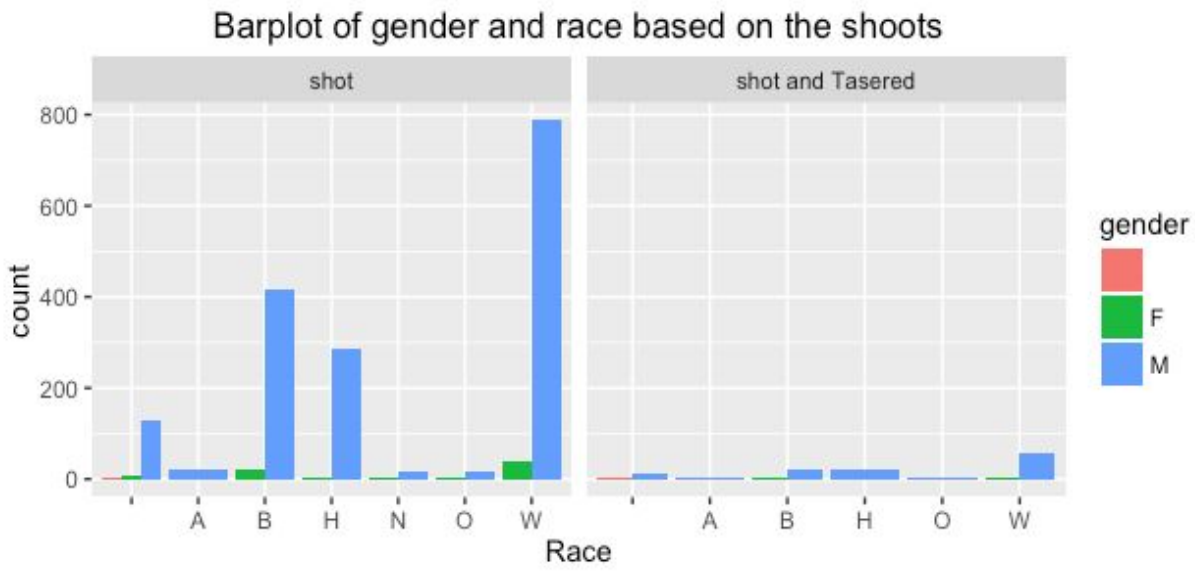
Also, there is an important relationship between the threat levels and manners of death. The graphs above illustrate that more of the victims have been perceived as attacks in terms of threat level, and that the most common method of death is being shot, no matter the threat level. The second most common of death is "other", which can include the possibility of non-threatening and non-attacking behavior. This can be interpreted to being unarmed unprovoked behavior. Lastly, undetermined is the least common method of death, which means that it is unclear whether the police officer was being attacked or not.

Race and Gender Analysis

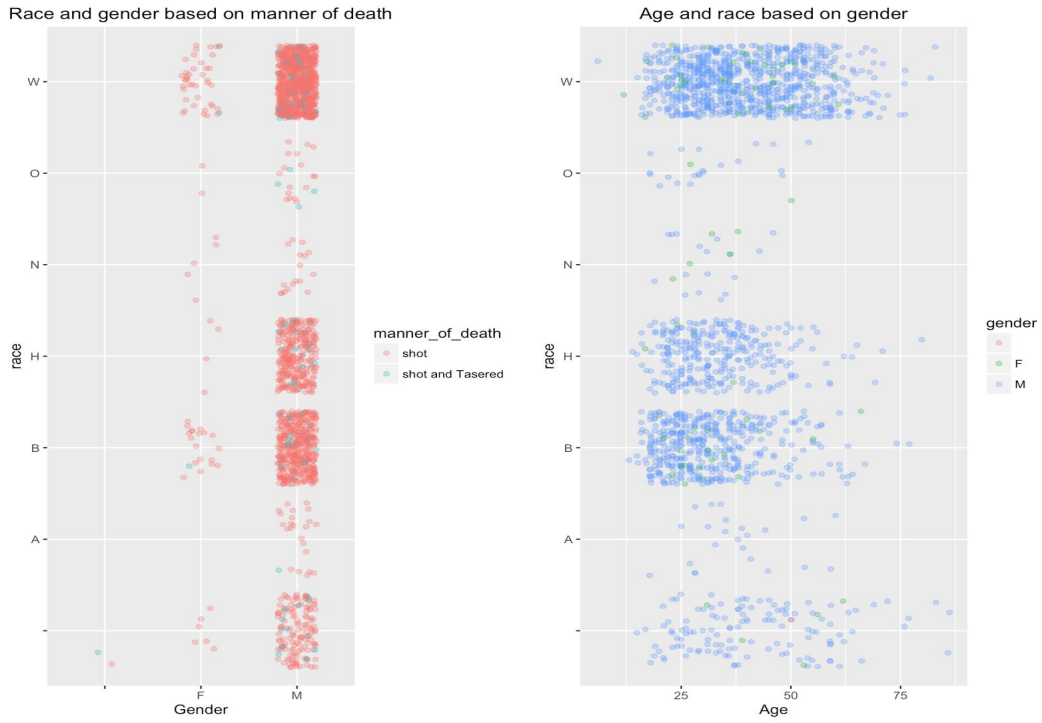
Since we are also interested in determining whether some of the races or genders are more likely to be shot, we use various plots to see their relationship. From the mosaic plot below, we see that regardless of the race most of the people are shot rather than shot and tasered. From the second plot, we see the frequency of each race. So, we see that there is a maximum number of White people followed by Black people and so on in this dataset.







From this bar plot, we can see that none of the Native American females were shot and tasered. Irrespective of the gender and number of people of each race, we see that there are very few Asian, Native American and other races that either get shot or shot and tasered.

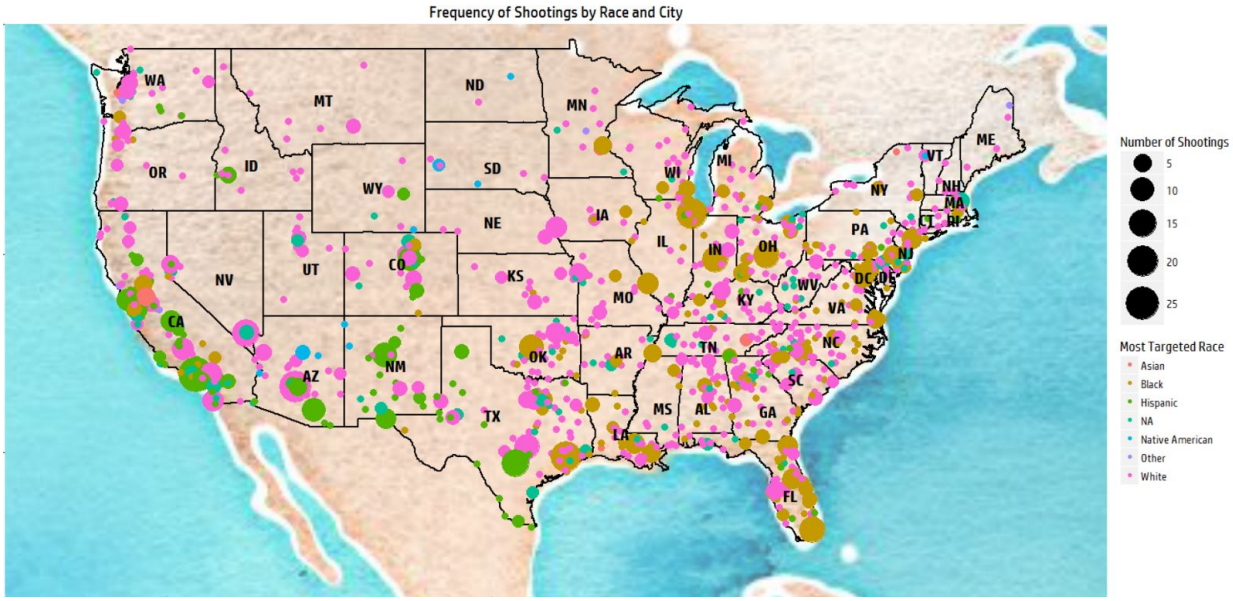


From the first graph, we can see that males are more likely to get shot as compared to females. We can also tell that most of the people that get shot are from three races: White, Black and Hispanic.

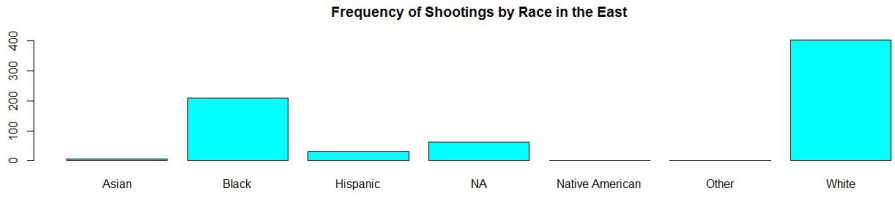
Also, most people die because they get shot rather than getting shot and tasered. From the second graph, we can see that most of the people that get shot are from the age group of 20 years to 50 years. However, the White race has a wider age for people getting shot. It also tells that there are more males than females as seen in the first graph.

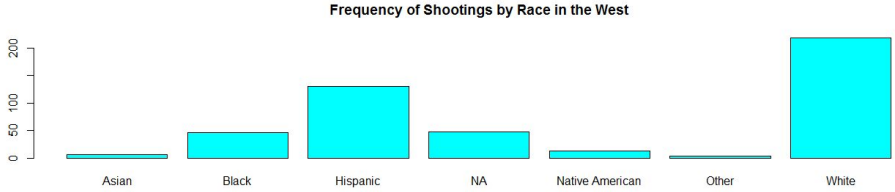
Map and Location Analysis

Finally, we are interested in analyzing the frequency and type of shooting based on location. The below map shows shootings by city. The size of each point is proportional to the number of shootings that occurred in that city. The color of the point depicts the most targeted race in that state. The map shows that, in general, whites are targeted the most.



Additionally, the most targeted races by city seem to shift as we move from east to west, with black highly targeted in the east and hispanics highly targeted in the west. Further, Native Americans are especially targeted in Arizona, New Mexico, Colorado, Wyoming, North Dakota and South Dakota. These findings are expected, given that there are more Native American reservations in these states. Other geographic However, these findings are not as significant without first standardizing the data.





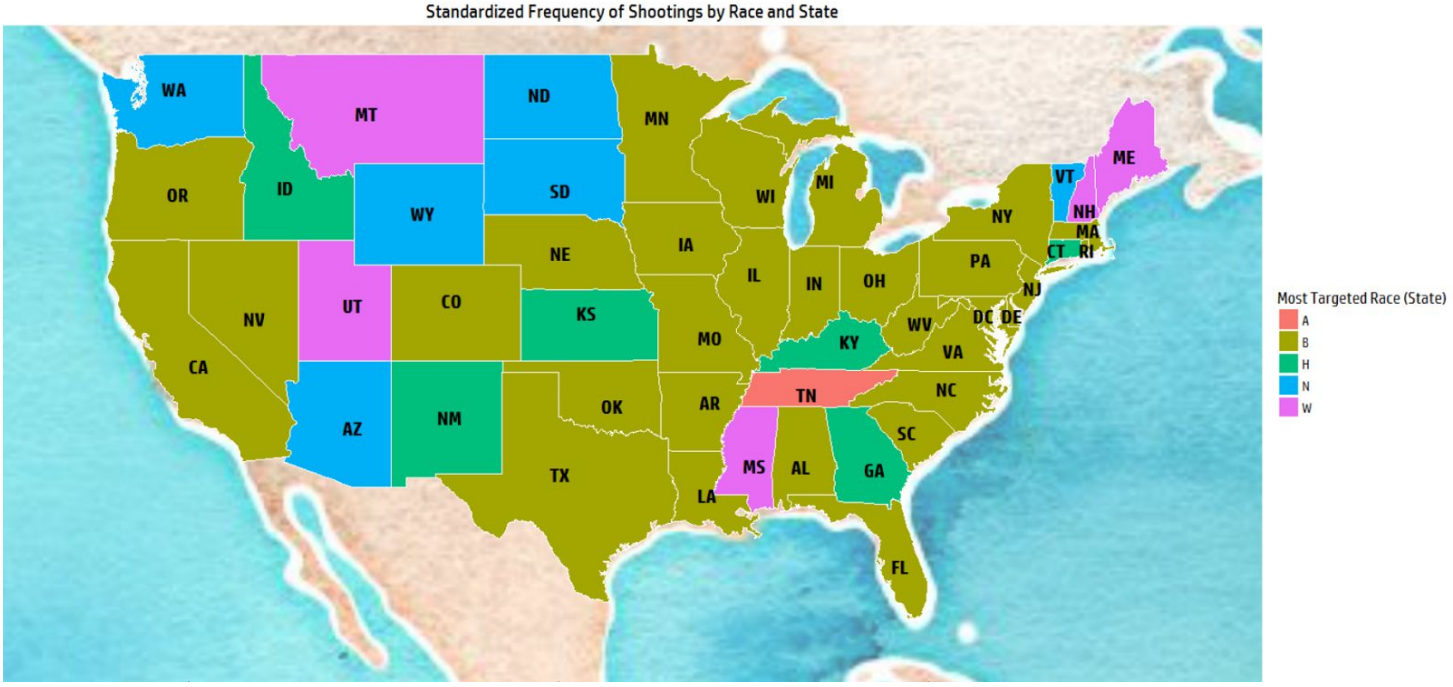
Here, NA refers to missing value - in other words, there was not a single race that qualified as the most targeted. Instead, two or more races were equally

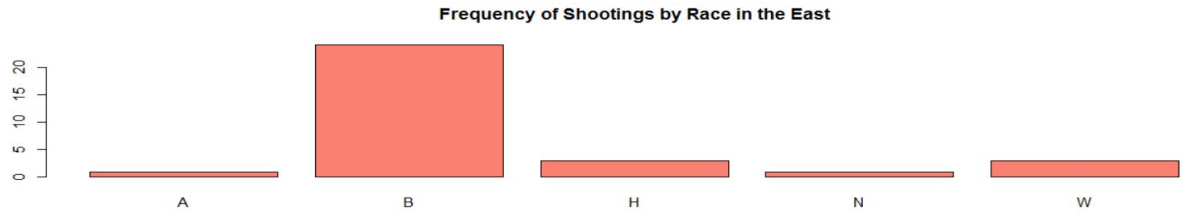
targeted in those areas.

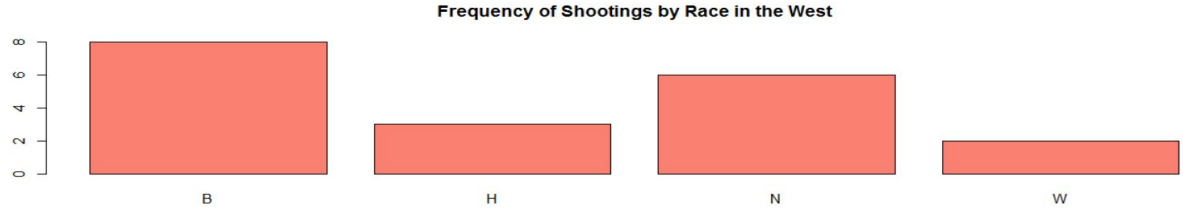
In understanding this data, one main point had to be addressed: whites comprise the majority of the population (around 70-80%). In other words, a police officer is *more likely to encounter* a white person, given their proportion of the population. Therefore, we must find whether racial groups are *disproportionately targeted* given their proportions in the general population.

To accomplish this, we sought additional census data from the US Census Bureau. This data gives the proportions of races in the population by state in 2008. While not a perfect representation of the current populations, the Census data provides an adequate estimate. For consistency with the Census data, Other and NA were removed when race standardization occurred, as these categories are not easily interpreted.

The below map provides a standardized representation of targeted races by state. Given their proportion of the population, blacks are the *most* targeted race in 33 states.







As before there is a clear difference between races targeted in the East and West. However, these standardized bar plots now show how minorities are disproportionately targeted.

IV. Conclusion

In exploring this data, we found several significant relationships among the variables. When analyzing the existence of a body camera, we found that police officers are much more likely to shoot when they are not wearing a body camera. In particular, young black and middle-aged white men were shot the most often, when police officers were not wearing a body camera. When a body camera was present, younger black men and older white men were the majority getting shot. Furthermore, we also discovered that most of the deaths were due to being perceived as attacks, and that shooting is the most common method of death. The next most common threat level was other, as in potentially non violent cases. Shooting is the most common method of death in all threat levels. In observing the frequency by which each race is shot, we found that white males are shot more than other races. Additionally, we found that hispanics are most targeted in the west and blacks are most targeted in the east. However, upon standardizing the data, we found that blacks, hispanics and native americans are disproportionately targeted. Whites and asians are targeted the least among all races. Once the data was standardized, it became clear that blacks are the most targeted, regardless of region.

V. Appendix: Code

Stephanie's Code:

```
library(ggplot2) #for creating plot
library(reshape) #for the function 'melt' to melt a dataframe for easy casting
library(mosaic) #for the mosaic plot to look for relationship
library(car) #for scatterplot Matrix
library(vcd)

shootingdata =
read.csv("https://raw.githubusercontent.com/washingtonpost/data-police-shootings/master/fatal-police
-shootings-data.csv")
head(shootingdata)
summary(shootingdata)
names(shootingdata)
attach(shootingdata)
armed.freq = table(armed)
sorted_armed <- sort(armed.freq, decreasing=T)
names(sorted_armed)
top5 = names(sorted_armed)[1:5]

shooting = subset(shootingdata,armed %in% top5)
detach(shootingdata)
shooting = droplevels(shooting)

levels(shooting$race) <- c("Unknown","A","B","H","N.A.","Other","W")
plot(shooting)
levels(shooting$flee) <-c("Unknown","Car","Foot","Not fleeing","Other")
levels(shooting$gender) <-c("Unknown","Female","Male")

attach(shooting)
summary(age)

pie(table(armed),
   main="Weapon Held by Civilians in Fatal Police Shooting\n (only 5 most common weapons
shown)", cex.main=1)
```

```
ggplot(data=shooting, aes(shooting$age)) +
  geom_histogram(aes(fill=..count..)) +
  xlim(c(0,90)) +
  ggtitle("Histogram of the distribution of age") +
  xlab("Age")+
  ylab("Frequency")

#Frequency of body camera
barplot(table(body_camera), main= "Frequency of body camera", xlab = "existence of body camera",
ylab = "Frequency",col = "blue")

#Barplot of existence of body camera and flee
ggplot(shooting, aes(body_camera, ..count..)) +
  geom_bar(aes(fill = flee), position = "dodge")+
  ggtitle("Barplot of existence of body camera and flee")+
  xlab("Body camera")


library(gridExtra)

graph1 = ggplot(shooting, aes(body_camera,flee, color = manner_of_death))+
  geom_jitter(position = position_jitter(width = 0.5) , alpha = 0.3) +
  ggtitle("Existence of body camera and flee\n based on manner of death") +
  xlab("Existence of body camera") +
  ylab("Flee")

graph2 = ggplot(shooting, aes(body_camera, threat_level, color = armed))+
  geom_jitter(position = position_jitter(width = 0.5) , alpha = 0.5) +
  ggtitle("Existence of body camera and threat level\n based on types of weapon") +
  xlab("Existence of body camera") +
  ylab("Threat level")

grid.arrange(graph1,graph2, ncol=2)

#Heatmap
racecounts.df <- as.data.frame(table(body_camera, race))
ggplot(racecounts.df, aes(body_camera, race)) +
  geom_tile(aes(fill = Freq), colour = "black") +
```

```
  scale_fill_gradient(low = "white", high = "steelblue")+
  ggtitle("heatmap of the frequency of existence of body camera\n at the time of shooting to different
race")+
  xlab("existence of body camera")

#Mosaic plots
shooting$age_cut = cut(shooting$age,3)
levels(shooting$age_cut) = c("young","middle-aged","older")

mosaic(~  race + body_camera | age_cut,
       data = shooting, shade = TRUE, legend = TRUE,main="Body camera and race based on different
age levels",direction = "v", rot_labels=c(0,90,0,0))

detach(shooting)
```

Dhanya's Code

```
#graphing threat level vs manner of death
 mod=factor(shootings$V4)
 tl=factor(shootings$V12)
counts<-table(mod,tl)
barplot(counts, main="Threat Level and Manner of Death", xlab="Threat Level", col=c("darkblue",
"red"), legend= rownames(counts))
 counts=counts[-1,-3]
barplot(counts, main="Threat Level and Manner of Death", xlab="Threat Level", col=c("darkblue",
"red"), legend= rownames(counts), beside=TRUE)
```

Janet's Code

```
ggplot(shootingdata, aes(race, ..count..)) + geom_bar(aes(fill = gender), position = "dodge")+
ggtitle("Barplot of gender and race based on the shoots")+ xlab("Race") +
+ facet_grid(.~manner_of_death, scale="free")



install.packages("gridExtra")
library(gridExtra)
```

```
plot1=ggplot(shootingdata, aes(gender,race, color = manner_of_death))+
geom_jitter(position = position_jitter(width = 0.5) , alpha = 0.3) +
ggtitle("Race and gender based on manner of death") +
xlab("Different Races")

plot2=ggplot(shootingdata, aes(gender,manner_of_death, color = race))+
geom_jitter(position = position_jitter(width = 0.5) , alpha = 0.3) +
ggtitle("Gender and manner of death based on race") +
xlab("Gender")

plot3=ggplot(shootingdata, aes(race,gender, color = manner_of_death))+
geom_jitter(position = position_jitter(width = 0.5) , alpha = 0.3) +
 ggtitle("race and gender based on death") +xlab("race")

plot4=ggplot(shootingdata, aes(age,race, color = gender))+
geom_jitter(position = position_jitter(width = 0.5) , alpha = 0.3) +
ggtitle("Age and race based on gender") +
xlab("Age")

grid.arrange(plot1, plot4, ncol=2)
grid.arrange(plot1, plot2, ncol=2)
grid.arrange(plot1, plot3, ncol=2)
grid.arrange(plot2, plot3, ncol=2)

table(shootingdata$gender, shootingdata$manner_of_death)
#shows Males have higher frequency

table(shootingdata$race, shootingdata$manner_of_death)
#This table shows 'Whites' have higher frequency

#Barplot of the frequency of race
ggplot(shootingdata, aes(race, ..count..)) +
  geom_bar(fill = rainbow(7))+
  ggtitle("Barplot of the frequency of race")+
```

```
  xlab("Race")
```

Juliana's Code:

```
#GGMAP CODE
#==========================
library(maps)
library(ggplot2)
library(ggmap)
library(data.table)
library(plyr)
library(extrafont)

#reading in the data
#==========================
shootings =
read.csv("https://raw.githubusercontent.com/washingtonpost/data-police-shootings/master/fatal-police
-shootings-data.csv")

#reading in Census Bureau Data
#============================
raceprop = read.csv("data/RACEPROPSTATE.csv")
raceprop = raceprop[,-c(2:7,13,14)]
colnames(raceprop) = c("State","W","B","N","A","H")

#Reading in location data about each city
#acquired by using apply geocode on city names, and exporting to csv
#=================================================================
shooting.loc = read.csv("data/GEOCODES.csv",header=FALSE)
newcity=paste0(city,", ", state, " United States")
cityf = count(newcity)
shootinglf = cbind(shooting.loc,cityf$freq)
names(shootinglf)= c("city","lon","lat","freq")


#data frame with race and city
```

```
#==============================
racecitydf = cbind(newcity,shootings$race)
racecity = table(racecitydf[,1], racecitydf[,2])

#find the most targeted race in each city
racemax = function(x){
  maxcol = c()
  a = which.max(x)
  f = names(x)[a]
  maxcol = append(maxcol,f)
  return(maxcol)
}
racemaxcol = apply(racecity,1,racemax)
racemaxcol = unname(racemaxcol)

rename = function(x){
  if(x == 1){
    x = 'NA'
  } else if(x==2){
    x = "Asian"
  } else if (x==3){
    x = "Black"
  } else if(x==4){
    x="Hispanic"
  } else if(x==5){
    x="Native American"
  } else if(x==6){
    x = "Other"
  } else if(x==7){
    x="White"
  }
}

racemaxcol = sapply(racemaxcol,rename)
racemaxcol = unname(racemaxcol)

shootinglf = cbind(shootinglf,racemaxcol)
```

```
#Finding Geocodes of States
#=========================================
colnames(shootinglf)=c("city","lon", "lat","Freq","racemaxcol")

b = levels(factor(state))
bstate = paste0(b, ", ", "United States")

statecodes = sapply(bstate, geocode)
s = t(statecodes)
row.names(s) = b

s[,1] = unname(unlist(s$lon))
s[,2]=unname(unlist(s$lat))
#didn't find oregon
oregon = geocode("Oregon, United States")
s[which(row.names(s)=="OR"),] = oregon
#centering NY more
s[which(row.names(s)=="NY"),] = c(-76,43)
s = as.data.frame(s)

#Frequency of shootings by Race and State
#=============================================
rsfreq = table(shootings$state, shootings$race)
rsfreq = rsfreq[,-c(1,6)]

#Standardizing race by state
rsfreq[,1] = rsfreq[,1]/raceprop$A
rsfreq[,2] = rsfreq[,2]/raceprop$B
rsfreq[,3] = rsfreq[,3]/raceprop$H
rsfreq[,4] = rsfreq[,4]/raceprop$N
rsfreq[,5] = rsfreq[,5]/raceprop$W

#Finding the mean standardized race frequency among all states
names(rsfreq)=c("Asian","Black","Hispanic","Native American","White")
rmcol = apply(rsfreq,1,racemax)

s = cbind(s, rmcol)
colnames(s)[3]="Max Race"
```

```
#Getting maps and map data
us = get_map("United States", zoom = 3, maptype="watercolor")
usa <- map_data("usa")
states <- map_data("state")

#Changing state data for use in ggmap
#need to make race a graphable ggmap aesthetic
states = states[,-6]

#changing states data to have matching state codes
for (i in 1:length(states$region)){

  if (states$region[i]=="alabama"){
    states$region[i]= "AL"
  } else if(states$region[i]=="alaska"){
    states$region[i] = "AK"
  } else if(states$region[i]=="arizona"){
    states$region[i] = "AZ"
  } else if(states$region[i]=="arkansas"){
    states$region[i] = "AR"
[...omitted]
  }else if(states$region[i]=="wisconsin"){
    states$region[i] = "WI"
  }else if(states$region[i]=="wyoming"){
    states$region[i] = "WY"
  }else if(states$region[i]=="district of columbia"){
    states$region[i] = "DC"
  }
}

findrace = function(){
  for (i in 1:length(states$region)){
    a = which(rownames(s)==states$region[i])
    states$maxrace[i] = as.character(s$`Max Race`[a])
  }
}
```

```
findrace()
#now states data has the most targeted race by state
#ggmap can now take most targeted race as a fill aesthetic


#Map 1: Frequencies by city
ggmap(us, darken = c(0.3, "white")) + borders("state",size=1,col="black") +
  geom_point(data = shootinglf,aes(x = lon, y = lat, size = Freq, color = racemaxcol)) +
  scale_size(range = c(3, 15))+
  geom_text(data = s,aes(x=lon,y=lat,label=row.names(s)),size=5,fontface="bold",
            family="HP Simplified", check_overlap = TRUE) +
  labs(size="Number of Shootings",color="Most Targeted Race") +
  ggtitle("Frequency of Shootings by Race and City") +
  xlab("Longitude") +
  ylab("Latitude") +
  theme(text=element_text(size=12, family="HP Simplified")) +
  coord_map(xlim = c(-130, -62),ylim = c(22, 50))

#subsetting east and west data
west = subset(shootinglf, subset = lon<(-96))
wcount = count(west$racemaxcol)
east = subset(shootinglf, subset = lon>-96)
ecount = count(east$racemaxcol)

#graphing east and west data
par(mfrow=c(2,1))
barplot(ecount$freq, names.arg = ecount$x, main = "Frequency of Shootings by Race in the East", col
= "cyan")
barplot(wcount$freq, names.arg = wcount$x, main = "Frequency of Shootings by Race in the West",
col = "cyan")


#Map 2: Standardized frequencies by state
ggmap(us, darken = c(0.3, "white")) +
  geom_polygon(data = states, aes(x = long, y = lat, fill = maxrace, group = group), color = "white") +
  geom_text(data = s,aes(x=lon,y=lat,label=row.names(s)),size=5,fontface="bold",
            family="HP Simplified", check_overlap = TRUE) +
```

```
  labs(fill="Most Targeted Race (State)") +
  ggtitle("Standardized Frequency of Shootings by Race and State") +
  xlab("Longitude") +
  ylab("Latitude") +
  theme(text=element_text(size=12, family="HP Simplified")) +
  coord_map(xlim = c(-130, -62),ylim = c(22, 50))

#graphing standardized max race data for whole country
rscount = count(s$`Max Race`)
barplot(rscount$freq, names.arg = rscount$x, main = "Standardized Frequency of Shootings by State
and Race", col = "green")

#subsetting standardized east and west data by state
wstates = subset(s, subset=lon<(-96))
wscount = count(wstates$`Max Race`)
estates = subset(s, subset=lon>-96)
escount = count(estates$`Max Race`)

#graphing standardized east and west data
par(mfrow=c(2,1))
barplot(escount$freq, names.arg = escount$x, main = "Frequency of Shootings by Race in the East",
col="salmon")
barplot(wscount$freq, names.arg = wscount$x, main = "Frequency of Shootings by Race in the West",
col="salmon")
```