Final Project

Alexa DiBenedetto and Danya Murali

4/9/2018

library(tidyverse)  
library(modelr)  
library(lubridate)  
library(waffle)  
library(gridExtra)  
library(splines)

dat <- read\_csv("~/Grad School /Data Science/Project/guns.csv")

#### Introduction

Gun data is notoriously hard to obtain, but given the current political climate and national conversation surrounding guns in America, it is more important than ever to collect and understand relevent data to inform future policy. Based on the limited amount of data made available by the CDC, we can begin to understand the circumstances surrounding gun crime in America. Our data covers gun deaths between 2012 and 2014, and collects information on various demographic and contextual factors. In total, there are just over 100,000 recorded gun-related deaths in this 3 year span.

Our main dataset was obtained from kaggle, and can be found at: <https://www.kaggle.com/hakabuk/gun-deaths-in-the-us/data> (1). The data originated from the CDC and the kaggle post was inspired by the Gun Deaths in America Project by FiveThirtyEight.com (<https://fivethirtyeight.com/features/gun-deaths/> (2)).

This project seeks to first explore the overall trends of gun crimes and the demographic breakdown of gun crime perpertrators and victims. Then we plan to replicate some of the interesting graphics found in FiveThirtyEight's article using tidyverse and related packages.

#### Variables and Metadata

The following variables are included in the dataset:

* Year - the year the gun death occured
* Month - the month the gun death occured
* Intent - the intent of the perpetrator
  + Homicide
  + Suicide
  + Accidental
  + Undetermined
  + NA
* Police - whether a police officer was involved in the shooting
* Sex - the sex of the victim
  + Male
  + Female
* Age - the age of the victim
* Race - the race of the victim
  + White
  + Black
  + Asian/Pacific Islander
  + Native American/Alaskan Native
  + Hispanic
* Hispanic - hispanic origin code
* Place - location of the shooting
  + Home
  + Farm
  + Industrial/construction
  + Residential Insititution
  + School
  + Sports
  + Street
  + Trade/service area
  + Other specified
  + Other unspecified
* Education - status of the victim
  + Less than high school
  + Graduated from high school
  + Some college
  + At least graduated from college
  + NA

#### Data Exploration

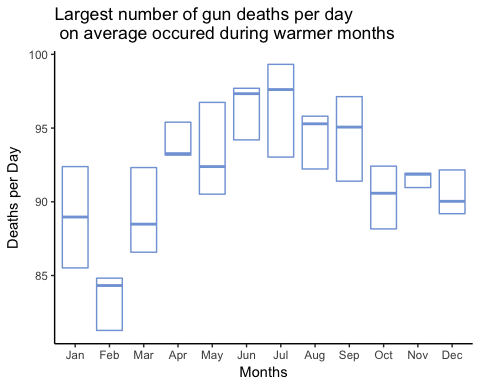
We began our data analysis by tidying up the original dataset to make it easier to extract relevant information. First, we created a month\_year variable that combines the month and year into a single date object. Next, we wanted to see the spread of gun deaths by month but realized that it was misleading in the current form of the data because months have differing days. To remedy this, we created a new variable called countPerDay, which computes the number of gun death per day by dividing the number of gun deaths per month by the number of days in the month.

Based on our boxplot of gun deaths per day by month, it can be determined that in general, there are more gun deaths in warmer months compared to colder months. The month with the least number of gun deaths per day on average was February, while the month with the most number of gun deaths per day on average was July.

dat <- dat %>% mutate(month\_year = paste(year, month, "01", sep="-")) %>%   
 mutate(month\_year = as.Date(month\_year)) %>% mutate(daysInMonth = days\_in\_month(month\_year))  
  
crimeCount <- dat %>% count(month\_year)   
crimeCount <- crimeCount %>% rename(count = n)   
  
  
crimeDat <- dat %>% left\_join(crimeCount) %>% mutate(countPerDay = count/daysInMonth) %>% mutate(education = as.factor(education)) %>% mutate(police = as.factor(police))

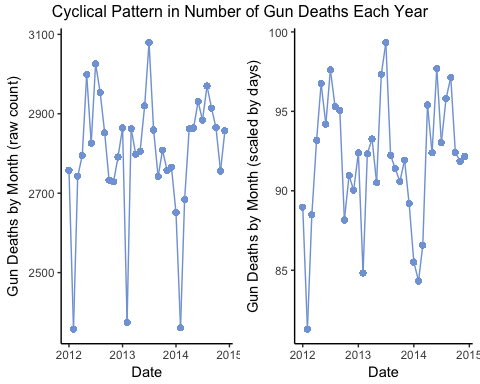
## Joining, by = "month\_year"

age\_cat <- function(age) {  
 cut(age,  
 breaks = c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110),  
 labels = c("0-10", "11-20", "21-30", "31-40", "41-50", "51-60", "61-70", "71-80", "81-90", "91-100", "101+")  
 )  
}  
  
crimeDat <- crimeDat %>% mutate(age\_cat = age\_cat(age)) %>%   
 mutate(month = as.numeric(month)) %>% mutate(month2 = month.abb[month])   
  
crimeDat %>% ggplot(aes(month2, countPerDay)) + geom\_boxplot(color="#81A3DA") +   
 ggtitle("Largest number of gun deaths per day \n on average occured during warmer months") +  
 xlab("Months") + ylab("Deaths per Day") + theme\_classic() + scale\_x\_discrete(limits = month.abb)



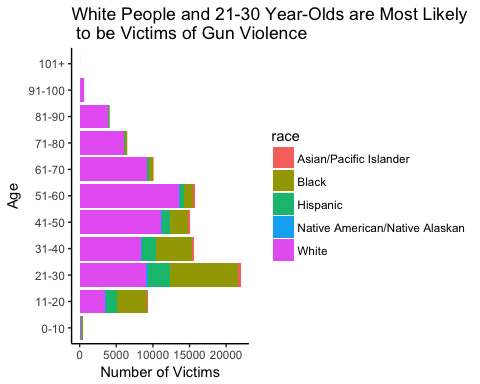
We plotted the number of gun deaths per month over the course of the three years in two ways: the raw count of gun deaths per month, and the adjusted count of gun deaths per month (scaled by number of days). We saw dramatic dips in Februrary in our first plot as a result of there being fewer days in February, however this phenomon still occured after we adjusted for the number of days in the month. Based on previous research, it is found that crime in general occurs less often when the temperature is colder.

# plot number of crimes per month  
plotmonth <- crimeDat %>% ggplot(aes(x = month\_year, y=count)) + geom\_point(color = "#81A3DA") +   
 geom\_line(color = "#81A3DA") + xlab("Date") + ylab("Gun Deaths by Month (raw count)") + theme\_classic()  
  
plotday <- crimeDat %>% ggplot(aes(x = month\_year, y=countPerDay)) + geom\_point(color = "#81A3DA") +   
 geom\_line(color = "#81A3DA") + xlab("Date") + ylab("Gun Deaths by Month (scaled by days)") + theme\_classic()  
  
grid.arrange(plotmonth, plotday, ncol=2, top="Cyclical Pattern in Number of Gun Deaths Each Year")



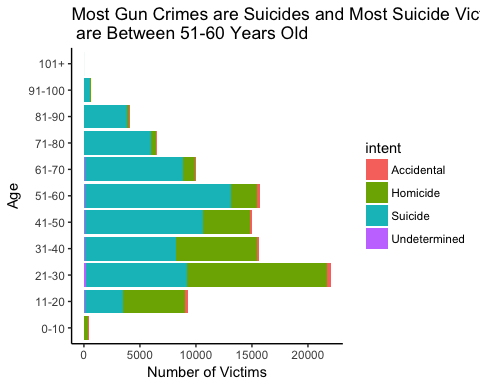
In comparing age to race, we found that black victims tend to be younger than white victims, and that most victims fall between the ages of 21-30. Also, we found that Asian/Pacific Islanders and Native American/Native Alaskans are the least likely to be victims of gun violence. White victims are most often between the ages of 51-60, while black and Hispanic victims are most often between 21-30.

crimeDat %>% na.omit(age\_cat) %>% ggplot(aes(x = age\_cat, fill=race)) + geom\_bar() +  
 coord\_flip() + ylab("Number of Victims") + xlab("Age") + theme\_classic() +  
 ggtitle("White People and 21-30 Year-Olds are Most Likely \n to be Victims of Gun Violence")



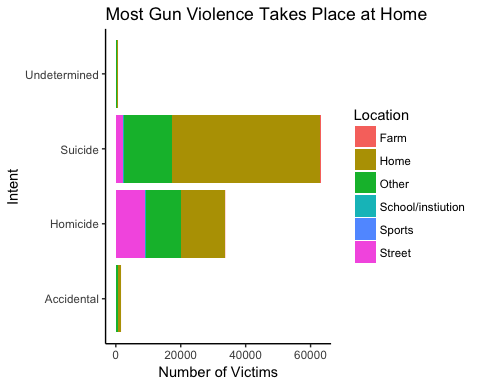
In comparing age to intent, we found that victims of homicide and accidental deaths are most often between the ages of 21-30, while victims of suicide are most often between the ages of 51-60. For victims between the ages of 0-10, deaths are only by homicide or accidental deaths, while for victims between the ages of 91-100, almost all deaths are by suicide.

crimeDat %>% na.omit(age\_cat) %>% ggplot(aes(x = age\_cat, fill = intent)) + geom\_bar() +  
 coord\_flip() + xlab("Age") + ylab("Number of Victims") + theme\_classic() +  
 ggtitle("Most Gun Crimes are Suicides and Most Suicide Victims \n are Between 51-60 Years Old")



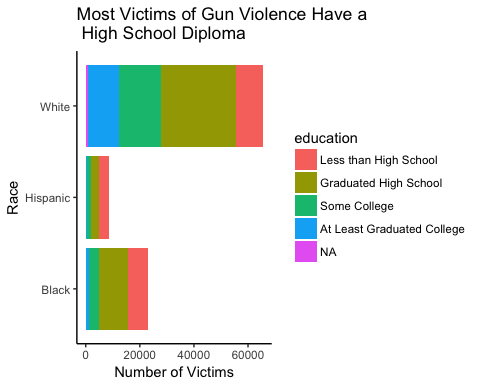
In comparing intent with location, we found that most suicides and homicides take place at home. Most gun crimes that occur on the streets are homicides.

crimeDat <- crimeDat %>%   
 mutate(Location = case\_when(place %in% c("Home", "Residential institution") ~ "Home",  
 place %in% c("Other specified", "Other unspecified",   
 "Industrial/construction", "Trade/service area") ~ "Other",  
 TRUE ~ as.character(place)))  
  
crimeDat %>% na.omit(age\_cat) %>% ggplot(aes(x = intent, fill = Location)) +   
 geom\_bar() + coord\_flip() + xlab("Intent") + ylab("Number of Victims") +  
 theme\_classic() +  
 ggtitle("Most Gun Violence Takes Place at Home")



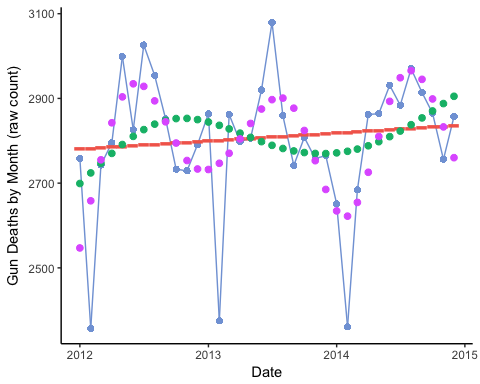
In comparing race and education of the victim, we found that for white and black people, most victims graduated from high school, whereas for hispanic people, most victims have less than a high school education.

crimeDat %>% filter(race %in% c("White", "Hispanic", "Black")) %>% na.omit(education) %>% ggplot(aes(x = race, fill = education)) +   
 geom\_bar() + coord\_flip() + xlab("Race") + ylab("Number of Victims") +   
 theme\_classic() + ggtitle("Most Victims of Gun Violence Have a \n High School Diploma") + scale\_fill\_discrete(labels=c("Less than High School", "Graduated High School",   
 "Some College", "At Least Graduated College", "NA"))



Since we only have the month and year as opposed to the day, we found it hard to create any meaningful model out of the data because we need month-level variables. As we can see from this plot, month vs. count does not have a linear relationship (red dotted line). We found that a natural spline with four degrees of freedom did a better job of describing the data and a natural spline with seven degrees of freedom did the best out of the three. However, all off these models still do not model the variation in gun crimes throughout the year well.

mod1 <- lm(count ~ month\_year, data = crimeDat)  
  
grid1 <- crimeDat %>%  
 data\_grid(month\_year) %>%  
 add\_predictions(mod1, "count")  
  
mod2 <- lm(count ~ ns(month\_year, df=4), data = crimeDat)  
  
grid2 <- crimeDat %>%  
 data\_grid(month\_year) %>%  
 add\_predictions(mod2, "count")  
  
mod3 <- lm(count ~ ns(month\_year, df=7), data = crimeDat)  
  
grid3 <- crimeDat %>%  
 data\_grid(month\_year) %>%  
 add\_predictions(mod3, "count")  
  
plotmonth + geom\_point(data = grid1, color = "#F4695A", size = 10, shape=45) +   
 geom\_point(data = grid2, color = "#00BA77", size = 2) +  
 geom\_point(data=grid3, color = "#E066FF", size = 2)



library(shiny)  
library(waffle)  
  
ui <- fluidPage(  
   
 titlePanel(  
 fluidRow(  
 column(4, offset = 4, "Gun Crime Data Visualization"))),  
   
 sidebarLayout(  
 sidebarPanel(  
 selectInput("variable",  
 "Please choose a variable:",  
 choices = c("Race", "Police Involvement", "Education", "Gender", "Intent", "Location of Crime")  
 ),  
   
 actionButton(inputId = "press", label = "Press for Plot")  
   
 ),  
   
   
 mainPanel(  
 tabsetPanel(  
   
 tabPanel("Plots", plotOutput("wafflePlot")),  
   
 tabPanel("Results", textOutput("waffleText"), tags$head(tags$style("#waffleText{font-size: 20px;}"))),  
   
 tabPanel("Description of the Data", tags$p("Gun data is hard to obtain. Based on the limited amount of data made available by the CDC, we can begin to understand the   
 circumstances surrounding gun crime in America. Our data covers gun deaths between 2012 and 2014, and collects data on various   
 demographic and contextual factors. In total, there were just over 100,000 recorded gun-related deaths in this 3 year span.   
 Data were collected from", tags$a(href="https://www.kaggle.com/hakabuk/gun-deaths-in-the-us/data", "Kaggle"), ".", style="font-size:20px"),  
 tags$ul(tags$li("Victim's Age", style="font-size:20px"),  
 tags$li("Victim's Gender", style="font-size:20px"),  
 tags$li("Victim's Race", style="font-size:20px"),  
 tags$li("Victim's Level of Education", style="font-size:20px"),  
 tags$li("Intent of Perpetrator", style="font-size:20px"),  
 tags$li("Month and Year of Crime", style="font-size:20px"),  
 tags$li("Location of Crime", style="font-size:20px"),  
 tags$li("Police Involvement", style="font-size:20px")))  
 )  
 )  
 )  
)  
  
server <- function(input, output) {  
  
 # race data  
 waffle\_race <- tribble(  
 ~White, ~Black, ~Asian, ~Hispanic, ~`Native American`,  
 66, 23, 1, 9, 1  
 )  
 # police data  
 waffle\_police <- tribble(  
 ~ Police, ~`No Police`,  
 1, 99  
 )  
 # education data  
 waffle\_edu <- tribble(  
 ~`Less than High School`, ~`High School`, ~`Some College`, ~`College`, ~`NA`,  
 22, 42, 22, 13, 1  
 )  
 # gender data  
 waffle\_gender <- tribble(  
 ~Female, ~Male,  
 14, 86  
 )  
 # intent data  
 waffle\_intent <- tribble(  
 ~Homocide, ~Suicide, ~Undetermined, ~Accidental,  
 35, 62, 1 , 2  
 )  
 # location data  
 waffle\_place <- tribble(  
 ~Home, ~Other, ~Street, ~School, ~Sports, ~Industrial, ~Farm,  
 61, 25, 10, 1, 1, 1, 1  
 )  
 # combined data  
 waffle\_data <- list(Race = waffle\_race, `Police Involvement` = waffle\_police, Education = waffle\_edu, Gender = waffle\_gender,  
 Intent = waffle\_intent, `Location of Crime` = waffle\_place)  
   
  
 button <- eventReactive(input$press, {  
 input$variable  
 })  
   
 waffle\_button <- eventReactive(input$press, {  
 waffle(unlist(waffle\_data[input$variable]))  
 })  
   
 text\_button <- eventReactive(input$press, {  
 if (input$variable == "Race"){  
 sentence <- "66% of gun crime victims are White. Only 23% are Black, 9% are Hispanic, 1% are Asian, and 1% are Native American"  
 } else if (input$variable == "Police Involvement"){  
 sentence <- "Only 1% of gun crimes involve a police officer."  
 } else if (input$variable == "Education"){  
 sentence <- "64% of gun crime victims have a high school diploma or less. 35% have a college degree or some college."  
 } else if (input$variable == "Gender"){  
 sentence <- "86% of gun crime victims are males."  
 } else if (input$variable == "Intent"){  
 sentence <- "62% of gun deaths are suicides. 35% are homocides, and 3% are accidental or undetermined."  
 } else if (input$variable == "Location of Crime"){  
 sentence <- "61% of gun deaths happen at home. 10% happen on the street, 1% happen in schools, and 27% happen in other locations."  
 }  
 })  
  
   
 # creates waffle plots  
 output$wafflePlot <- renderPlot({  
 waffle\_button()  
 })  
  
 # generates accompanying text for the plots  
 output$waffleText <- renderText({  
 text\_button()  
 })  
}  
  
# Run the application   
shinyApp(ui = ui, server = server)

#### Discussion

Through this analysis we found that gun crime affects populations and demographics in different ways. In general, we found that more gun deaths occur in the warmer months than the colder months. This is mirrored in general crime statistics, where on average there is an uptick in crime when the weather is warm (3).

In comparing race and age, we found that black victims tended to be younger than white victims. White victims are most often between the ages of 51-60, while black and hispanic victims are most often between the ages of 21-30. Overall, most victims fall between the ages of 21-30. Asian/Pacific Islanders and Native Americans/Native Alaskans are the least likely to be victims of gun violence.

In comparing intent and age, we found that victims of homicide and accidental death are most often between the ages of 21-30. Victims of suicide are most often between the ages of 51-60. For victims younger than 10, all gun deaths were homicides or accidental, while for victims older than 91, almost all gun deaths are by suicide.

In comparing intent and location, we found that most suicides and homicides take place at home. Of the gun crimes that occur on the street, most are homicides.

In comparing race and education, we found that for black and white people, most victims have a high school degree, while most hispanic victims have less than a high school degree.

In terms of modeling gun deaths by date, we found that there is no clear linear relationship. Next we tried a spline with four degrees of freedom which still did not account for enough of the seasonal variation in the data. Finally, we tried a spline with seven degrees of freedom which did a better job of capturing the seasonal variation of the data, but is still not an ideal model because it leaves out many important factors. Because our data is aggregated monthly and we don't have month-level predictors in this dataset, our regression analysis is limited in this project. In the future, we might look to add month-level predictors such as average temperature, political factors, economic factors etc. to improve the predictive power of this model.

In conclusion, this type of data analysis is crucial to making informed policy involving guns in America.

#### References

1. <https://www.kaggle.com/hakabuk/gun-deaths-in-the-us/data>
2. <https://fivethirtyeight.com/features/gun-deaths/>
3. <http://drexel.edu/now/archive/2017/September/Violent-Crime-Increases-During-Warmer-Weather-No-Matter-the-Season/>
4. <https://alexa-dibenedetto.shinyapps.io/GunDeaths/> (Link to Shiny App)