AprilM\_11 Final Project

April Meyer

2/21/2020

#### Introduction

After the 2014 killing of Michael Brown in Ferguson, Missouri, The Washington Post has started to collect a database of every fatal shooting in the US by a police officer in the line of duty. The protest movement, Black Lives Matter has increased focus on police accountability. This data includes information about the deceased including race, age, gender, if they were armed, mental health, fleeing and location. The post does not track deaths of people in police custody, non-shooting deaths, or by off-duty police. They aim to document conditions most like the Michael Brown in Ferguson. Centers for Disease control and FBI track fatal shootings by police, but according to Washington Post their data is incomplete. There has not been much dependable data around shootings by police before this movement and I would like to understand the current trends among fatal police shootings.

#### The Problem Statement

After the fatal shooting in 2014 waves of public protest broke out and lead to the movement of Black Lives Matter. I aim to look at the Washington Post data to see if there are trends that show signs of biases.

#### Analysis

My analysis included importing and cleaning the data. Once that was complete I sliced the data sets to compare variables.

#importing into a data frame  
PoliceKillingsUS <- read.csv("PoliceKillingsUS.csv")  
#Starting to clean and manipulate data  
# Replace all empty strings in flee & race with NA  
PoliceKillingsUS$flee[PoliceKillingsUS$flee == ""] <- NA  
PoliceKillingsUS$race[PoliceKillingsUS$race == ""] <- NA  
PoliceKillingsUS$armed[PoliceKillingsUS$armed == ""] <- NA  
#remove all rows with any missing values  
PoliceKillingsUS\_clean <- na.omit(PoliceKillingsUS)  
# Apply separate() to date  
PoliceKillingsUS\_clean <- separate(PoliceKillingsUS\_clean, col = date, into = c("Day", "Month", "Year" ), sep = "/")  
PoliceKillingsUS\_clean$Month <- as.factor(PoliceKillingsUS\_clean$Month)  
PoliceKillingsUS\_clean$Year <- as.factor(PoliceKillingsUS\_clean$Year)  
#cleaning up name column by converting to character and trimming  
PoliceKillingsUS\_clean$name <- as.character(PoliceKillingsUS\_clean$name)  
PoliceKillingsUS\_clean$name <- str\_trim(PoliceKillingsUS\_clean$name)   
#cleaning up city & state column by converting to character and trimming  
PoliceKillingsUS\_clean$name <- as.character(PoliceKillingsUS\_clean$city)  
PoliceKillingsUS\_clean$name <- str\_trim(PoliceKillingsUS\_clean$city)   
PoliceKillingsUS\_clean$name <- as.character(PoliceKillingsUS\_clean$state)  
PoliceKillingsUS\_clean$name <- str\_trim(PoliceKillingsUS\_clean$state)   
# Apply unite() to PoliceKillingsUS\_clean  
PoliceKillingsUS\_cc <- unite(PoliceKillingsUS\_clean, city\_st, city, state, sep = ", ")  
PoliceKillingsUS\_cc$city\_st <- as.factor(PoliceKillingsUS\_cc$city\_st)  
# Check the structure  
str(PoliceKillingsUS\_cc)

## 'data.frame': 2254 obs. of 15 variables:  
## $ id : int 3 4 5 8 9 11 13 15 16 17 ...  
## $ name : chr "WA" "OR" "KS" "CA" ...  
## $ Day : chr "02" "02" "03" "04" ...  
## $ Month : Factor w/ 12 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Year : Factor w/ 3 levels "15","16","17": 1 1 1 1 1 1 1 1 1 1 ...  
## $ manner\_of\_death : Factor w/ 2 levels "shot","shot and Tasered": 1 1 2 1 1 1 1 1 1 1 ...  
## $ armed : Factor w/ 69 levels "","air conditioner",..: 26 26 66 65 45 26 26 26 66 65 ...  
## $ age : int 53 47 23 32 39 18 22 35 34 47 ...  
## $ gender : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 1 2 ...  
## $ race : Factor w/ 7 levels "","A","B","H",..: 2 7 4 7 4 7 4 7 7 3 ...  
## $ city\_st : Factor w/ 1382 levels "Abbeville, AL",..: 1134 23 1345 1093 395 514 221 56 172 642 ...  
## $ signs\_of\_mental\_illness: logi TRUE FALSE FALSE TRUE FALSE FALSE ...  
## $ threat\_level : Factor w/ 3 levels "attack","other",..: 1 1 2 1 1 1 1 1 2 1 ...  
## $ flee : Factor w/ 5 levels "","Car","Foot",..: 4 4 4 4 4 4 2 4 4 4 ...  
## $ body\_camera : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

# View a summary  
summary(PoliceKillingsUS\_cc)

## id name Day Month Year   
## Min. : 3.0 Length:2254 Length:2254 02 :251 15:947   
## 1st Qu.: 705.2 Class :character Class :character 01 :244 16:873   
## Median :1346.0 Mode :character Mode :character 03 :244 17:434   
## Mean :1362.0 04 :206   
## 3rd Qu.:2002.8 06 :200   
## Max. :2820.0 05 :193   
## (Other):916   
## manner\_of\_death armed age gender race   
## shot :2100 gun :1247 Min. : 6.00 F: 94 : 0   
## shot and Tasered: 154 knife : 324 1st Qu.:26.00 M:2160 A: 36   
## unarmed : 165 Median :34.00 B: 592   
## vehicle : 158 Mean :36.25 H: 401   
## toy weapon : 102 3rd Qu.:45.00 N: 29   
## undetermined: 92 Max. :91.00 O: 28   
## (Other) : 166 W:1168   
## city\_st signs\_of\_mental\_illness threat\_level   
## Los Angeles, CA: 35 Mode :logical attack :1462   
## Phoenix, AZ : 28 FALSE:1681 other : 677   
## Houston, TX : 23 TRUE :573 undetermined: 115   
## Chicago, IL : 22   
## Las Vegas, NV : 17   
## Austin, TX : 16   
## (Other) :2113   
## flee body\_camera   
## : 0 Mode :logical   
## Car : 360 FALSE:2002   
## Foot : 278 TRUE :252   
## Not fleeing:1528   
## Other : 88   
##   
##

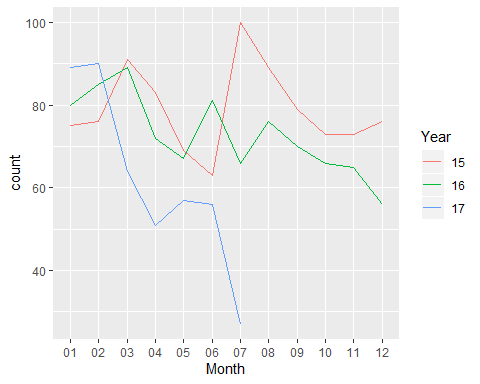
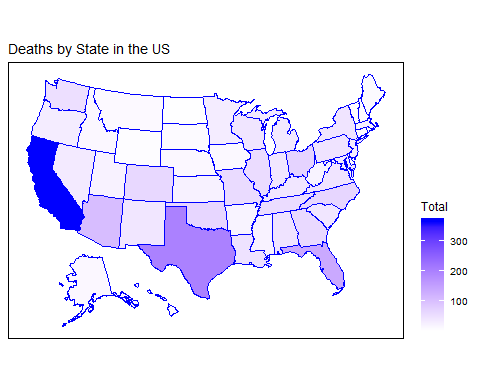
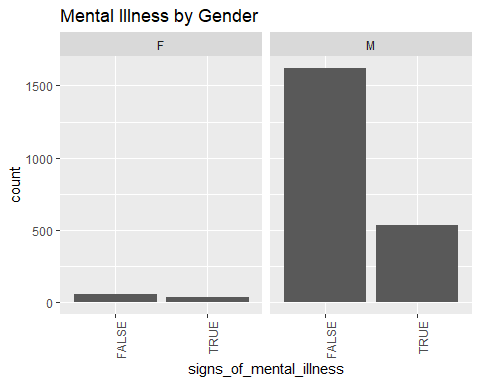
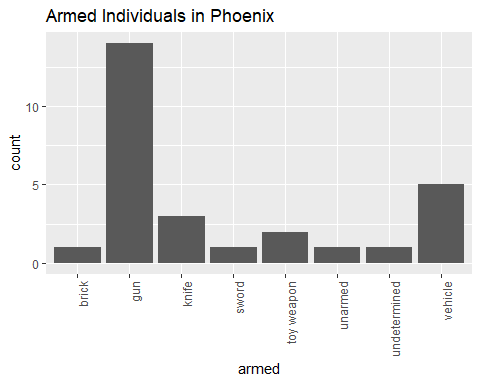
#creating different variables female and male  
femaleSubset <- subset(PoliceKillingsUS\_cc, subset = gender == "F")  
maleSubset <- subset(PoliceKillingsUS\_cc, subset = gender == "M")  
  
#creating different variables bodyCamera  
noCameraSubset <- subset(PoliceKillingsUS\_cc, subset = body\_camera == FALSE)  
CameraSubset <- subset(PoliceKillingsUS\_cc, subset = body\_camera == TRUE)  
  
#creating different variables mental illness  
noSignsMentalSubset <- subset(PoliceKillingsUS\_cc, subset = signs\_of\_mental\_illness == FALSE)  
SignsMentalSubset <- subset(PoliceKillingsUS\_cc, subset = signs\_of\_mental\_illness == TRUE)  
  
#creating different variables mental illness  
LosAngelesSubset <- subset(PoliceKillingsUS\_cc, subset = city\_st == "Los Angeles, CA")  
PhoenixSubset <- subset(PoliceKillingsUS\_cc, subset = city\_st == "Phoenix, AZ")

#### Implications

I have done some plots to show how variables compare and totals for some. I have also attempted to fit a glm model to see if a person is armed. Age and signs of mental illness would be worth looking into further. Since they have the greatest effect because the Pr(>|z|) is less than 0.05 for those four variables.

## # A tibble: 67 x 3  
## `PoliceKillingsUS\_cc$armed` Count Percent  
## <fct> <int> <dbl>  
## 1 gun 1247 55  
## 2 knife 324 14  
## 3 unarmed 165 7  
## 4 vehicle 158 7  
## 5 toy weapon 102 5  
## 6 undetermined 92 4  
## 7 machete 16 1  
## 8 unknown weapon 15 1  
## 9 ax 8 0  
## 10 sword 8 0  
## # ... with 57 more rows

## # A tibble: 1,382 x 3  
## `PoliceKillingsUS\_cc$city\_st` Count Percent  
## <fct> <int> <dbl>  
## 1 Los Angeles, CA 35 2  
## 2 Phoenix, AZ 28 1  
## 3 Houston, TX 23 1  
## 4 Chicago, IL 22 1  
## 5 Las Vegas, NV 17 1  
## 6 Austin, TX 16 1  
## 7 San Antonio, TX 15 1  
## 8 Columbus, OH 14 1  
## 9 Indianapolis, IN 14 1  
## 10 Miami, FL 14 1  
## # ... with 1,372 more rows



## 'data.frame': 2254 obs. of 6 variables:  
## $ armed : int 1 1 0 1 1 1 1 1 0 1 ...  
## $ age : int 53 47 23 32 39 18 22 35 34 47 ...  
## $ gender : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 1 2 ...  
## $ race : Factor w/ 7 levels "","A","B","H",..: 2 7 4 7 4 7 4 7 7 3 ...  
## $ signs\_of\_mental\_illness: logi TRUE FALSE FALSE TRUE FALSE FALSE ...  
## $ flee : Factor w/ 5 levels "","Car","Foot",..: 4 4 4 4 4 4 2 4 4 4 ...  
## - attr(\*, "na.action")= 'omit' Named int 60 125 242 267 341 399 418 427 471 519 ...  
## ..- attr(\*, "names")= chr "60" "125" "242" "267" ...

##   
## Call:  
## glm(formula = armed ~ ., family = binomial(), data = PoliceKillingsUS\_armed)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5682 0.3383 0.4475 0.5493 0.8547   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.500802 1.104111 2.265 0.0235 \*   
## age 0.029237 0.006412 4.559 5.13e-06 \*\*\*  
## genderM -0.028000 0.364700 -0.077 0.9388   
## raceB -1.632691 1.024628 -1.593 0.1111   
## raceH -1.768604 1.027429 -1.721 0.0852 .   
## raceN -0.828540 1.256452 -0.659 0.5096   
## raceO -1.703249 1.155488 -1.474 0.1405   
## raceW -1.332913 1.023706 -1.302 0.1929   
## signs\_of\_mental\_illnessTRUE 0.428516 0.183574 2.334 0.0196 \*   
## fleeFoot -0.294580 0.229945 -1.281 0.2002   
## fleeNot fleeing 0.054598 0.186300 0.293 0.7695   
## fleeOther -0.068175 0.352410 -0.193 0.8466   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1599.6 on 2253 degrees of freedom  
## Residual deviance: 1534.4 on 2242 degrees of freedom  
## AIC: 1558.4  
##   
## Number of Fisher Scoring iterations: 6

#### Limitations

There were many limitations to this data set. First once I started digging into the data more there were many undefined, missing, and NA values. Also, there were inaccuracies in data due to duplicates. The source of the data caused problems; for example, using only existing newspaper sources which could not include small towns or shootings that did not make the newspapers. Lastly it was harder to do any predictive modelling due to all the categorical data.  
#### Concluding Remarks  
I enjoyed working on this project; however, I do wish I would have picked a better data set. I did not realize how many limitations I would encounter with this set. Furthermore, I want to learn more on how to prepare and clean data. That would have been helpful with this data set.