Math 130 Project

Jason Ellis

2023-09-15

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=50),tidy=TRUE)
```

```
library(readxl)
P_shootings <- read_excel("C:/Users/jason/Desktop/Math 130/Data/fatal-police-shootings-data.xlsx",
    sheet = 1, col_names = TRUE)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Math 130 Project

Introduction

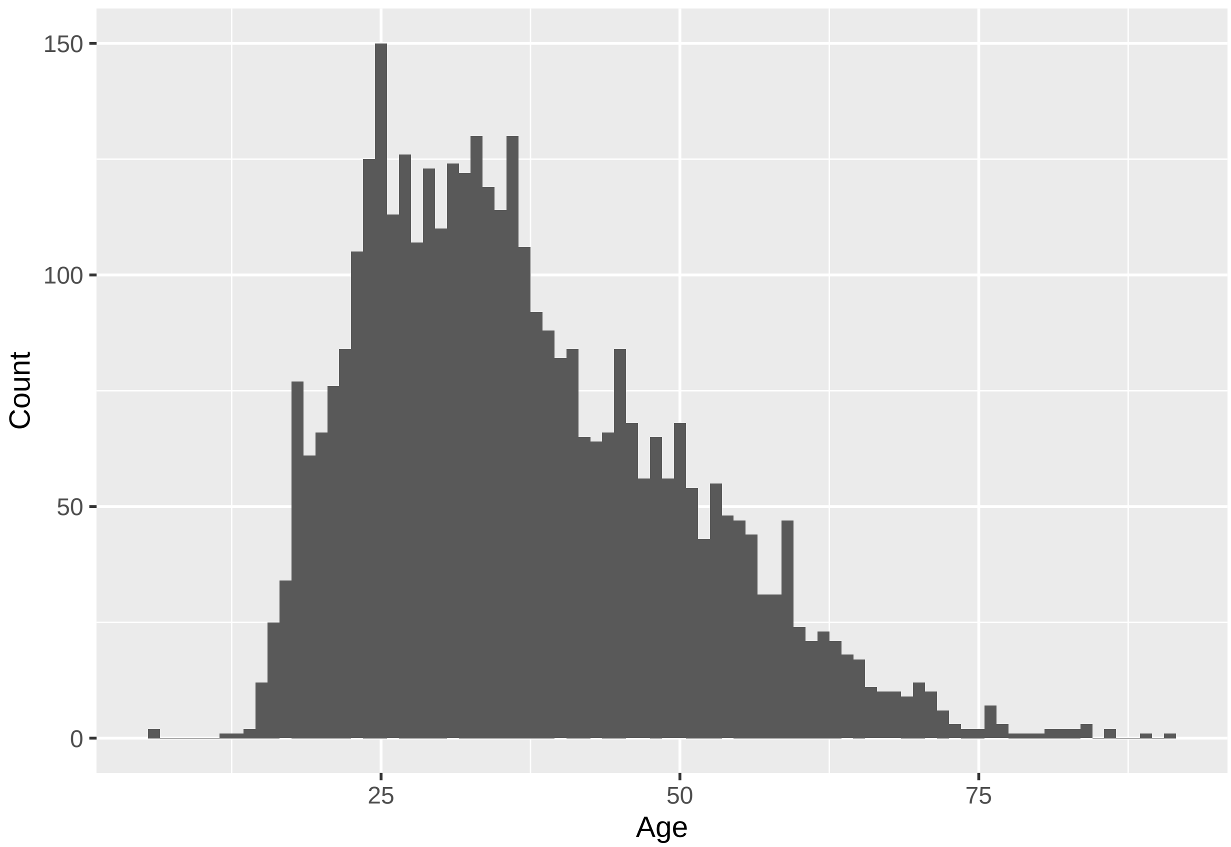
This data set is a collection of fatal police shootings that occurred in 2015. The data only includes shootings done by an on duty officer who shoots a civilian. This data does not include off duty officers, people already in police custody, or civilians who are killed from non shooting related circumstances. This data can be referenced on the website: https://github.com/washingtonpost/data-police-shootings

For my exploratory data analysis I will be comparing the variables of age and gender. My hypothesis is that at least 50% of these police shootings will have occurred upon males aged 18-30. I chose this as my hypothesis because there is a societal view and a common belief that younger males break the law more often. Due to this, I am interested to see if this anecdotal evidence is reflected in the number of fatal police shootings during the year of 2015.

Univariate Exploration

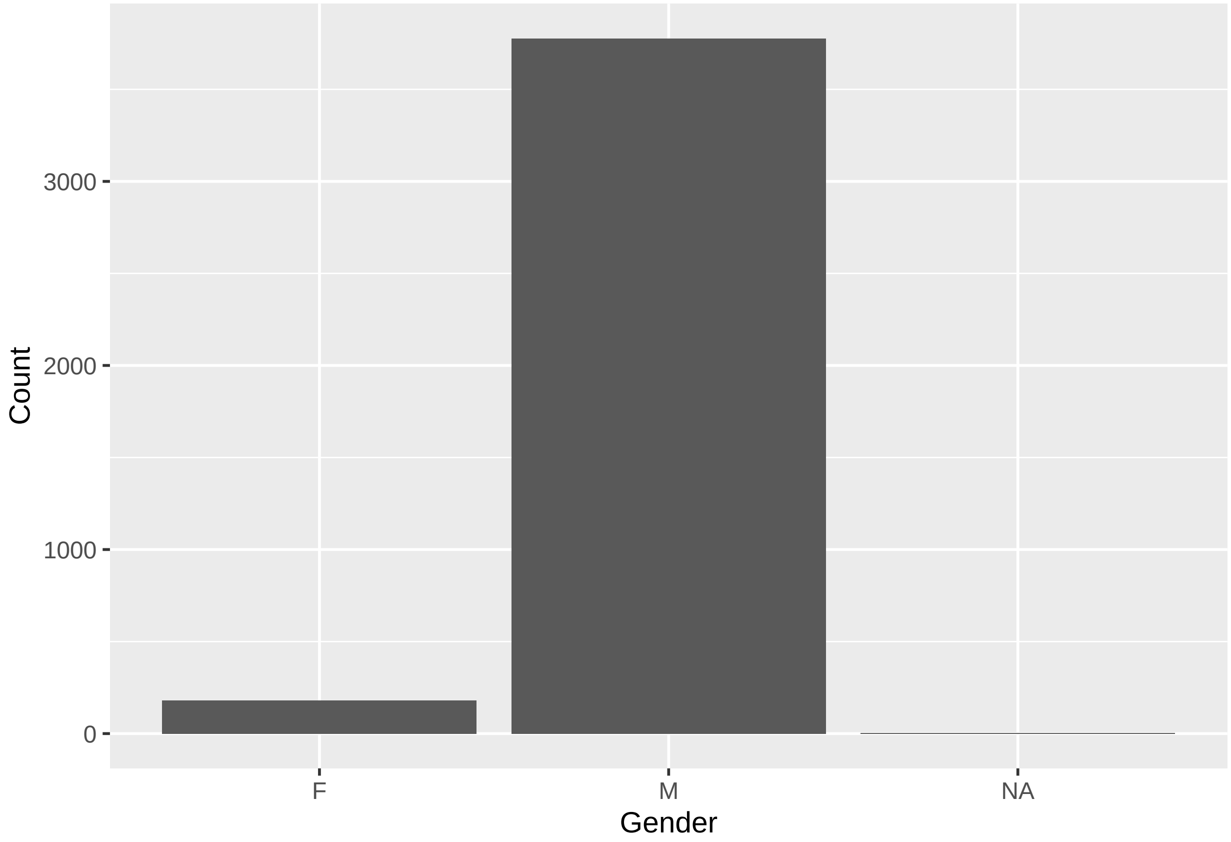
Before we are able to compare age and gender together we need to look at each variable individually. First, we can observe the age data. A histogram will do a good job of displaying this data.

```
ggplot(P_shootings, aes(age)) + geom_histogram(bins = 86) +
    xlab("Age") + ylab("Count")
```



This histogram does not separate the genders, but rather shows their data combined. The graph shows how a large percentage of deaths of both males and females occurred between the age of 25 and 35. Next we can take a look at the gender variable. The best way of displaying this data will be with a bar graph.

```
ggplot(P_shootings, aes(gender)) + geom_bar() + xlab("Gender") +
    ylab("Count")
```

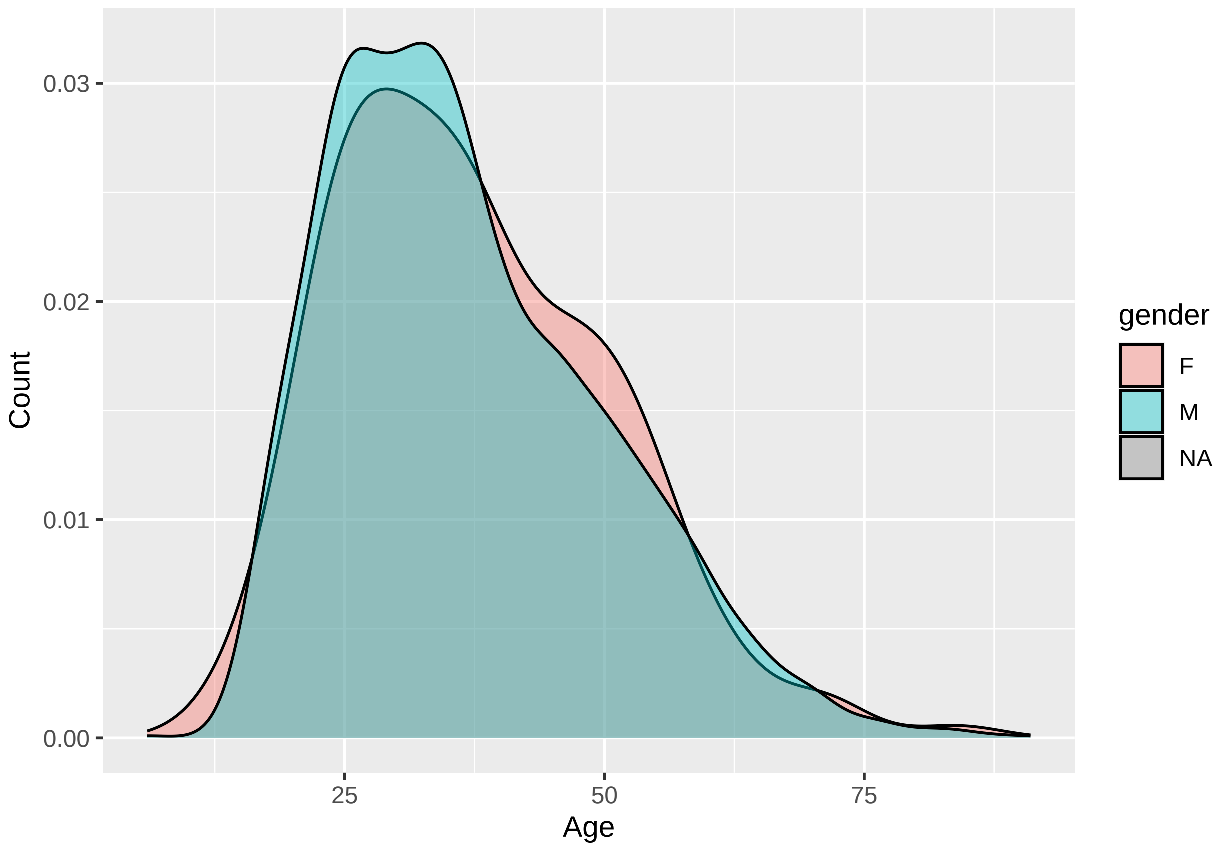


The bar graph clearly shows how the vast majority of deaths in this data set are males.

Bivariate Exploration

Now we can compare both of the variables. A density plot will show the distribution of the proportion of males and females.

```r
ggplot(P_shootings, aes(x = age, fill = gender)) +
    geom_density(alpha = 0.4) + xlab("Age") + ylab("Count")
```



This density plot shows how a large proportion of the deaths in this data set whether male or female occur between the age of 25 and 35. It simply shows the percentage of males and females separately, and not total deaths. A table will be a good tool to see how many deaths occur to each gander and how old they were at the time.

```r
table(P_shootings$gender, P_shootings$age, useNA = "always") %>%
    prop.table() %>%
    round(4) * 100
```

```
##
##           6   12   13   14   15   16   17   18   19   20   21   22   23   24
##   F    0.00 0.03 0.00 0.00 0.00 0.03 0.10 0.03 0.00 0.05 0.13 0.05 0.15 0.13
##   M    0.05 0.00 0.03 0.05 0.30 0.61 0.76 1.92 1.54 1.62 1.77 2.07 2.50 3.03
##   <NA> 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.03 0.00 0.00 0.00
##
##          25   26   27   28   29   30   31   32   33   34   35   36   37   38
##   F    0.23 0.13 0.15 0.13 0.05 0.15 0.20 0.10 0.10 0.15 0.05 0.18 0.15 0.13
##   M    3.56 2.73 3.03 2.58 3.06 2.63 2.93 2.98 3.18 2.85 2.83 3.11 2.53 2.20
##   <NA> 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##
##          39   40   41   42   43   44   45   46   47   48   49   50   51   52
##   F    0.13 0.10 0.08 0.05 0.08 0.08 0.08 0.15 0.03 0.08 0.13 0.15 0.03 0.00
##   M    2.10 1.97 2.05 1.59 1.54 1.59 2.05 1.57 1.39 1.57 1.29 1.57 1.34 1.09
##   <NA> 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##
##          53   54   55   56   57   58   59   60   61   62   63   64   65   66
##   F    0.10 0.13 0.10 0.05 0.00 0.00 0.03 0.05 0.03 0.05 0.00 0.00 0.00 0.03
##   M    1.29 1.09 1.09 1.06 0.78 0.78 1.16 0.56 0.51 0.53 0.53 0.45 0.43 0.25
##   <NA> 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##
##          67   68   69   70   71   72   73   74   75   76   77   78   79   80
##   F    0.00 0.00 0.00 0.00 0.05 0.03 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##   M    0.25 0.25 0.23 0.30 0.20 0.13 0.08 0.05 0.05 0.18 0.08 0.03 0.03 0.03
##   <NA> 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##
##          81   82   83   84   86   89   91 <NA>
##   F    0.00 0.00 0.00 0.03 0.00 0.00 0.00 0.18
##   M    0.05 0.05 0.05 0.05 0.05 0.03 0.03 3.61
##   <NA> 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.05
```

The histogram and density plot showed how the data was spread out but didn't show specific numbers. This table shows the breakdown of each gender and how many deaths occurred at each age. This table is a bit much to look at so I am going to manipulate the age variable so that they are placed neatly into groups of 12 by creating a new variable called "age_range". Once this is done it will create the 18-30 age group we are attempting to observe.

```
P_shootings$age_range <- cut_width(P_shootings$age,
    width = 12)
```

```
table(P_shootings$gender, P_shootings$age_range, useNA = "always") %>%
    prop.table() %>%
    round(4) * 100
```

```
##
##        [6,18] (18,30] (30,42] (42,54] (54,66] (66,78] (78,90] (90,102]  <NA>
##   F      0.18    1.34    1.41    1.01    0.33    0.08    0.03     0.00  0.18
##   M      3.71   30.10   30.30   17.35    8.13    1.82    0.33     0.03  3.61
##   <NA>   0.00    0.03    0.00    0.00    0.00    0.00    0.00     0.00  0.05
```

This proportion table is much easier to look at and neatly places the genders into age categories. It is easy to read shows that 30.1% of deaths are males of age 18-30.

Conclusion

In conclusion, my hypothesis cannot be supported by this data. My hypothesis of at least 50% of these police shootings will have occurred upon males aged 18-30 is close but not within the parameters for success. The data shows that males age 18-30 only accounted for 30.1% of deaths in this data set.