Research Project

Isabella Macayan

2023-09-15

```
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
```

```
library(RColorBrewer)
library(ggplot2)
```

Introduction

The data that I will be analyzing in this report is the police shootings data set, which provides an excel file on police shootings in 2015 as compiled by the Washington Post. It consists of 3960 observations and 14 variables. The variables I will be focusing on is the age and whether they displayed signs of mental illness or not. My research question is if younger people will more often get killed by the police than older people if they display signs of mental illness.

```
library(readxl)

police <- read_excel("/Users/izzymacayan/Desktop/math130/data/fatal-police-shootings-data.xlsx")
dim(police)
```

```
## [1] 3960   14
```

Univariate Exploration

Variable: Age

Summary Statistics

The information below shows the summary statistics for the continuous numerical variable, age. The mean/average age for police-involved killings is 36.85.

```r
summary(police$age)
```

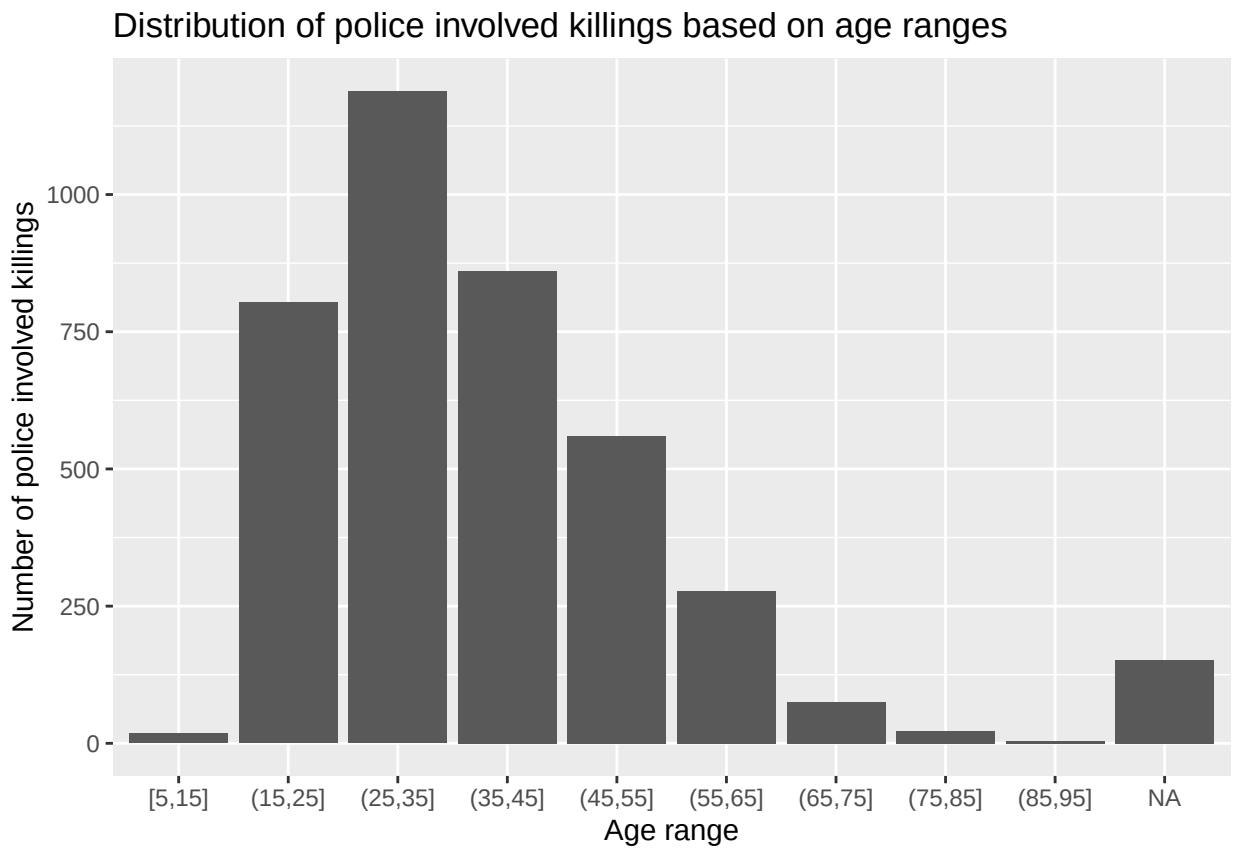
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   27.00   35.00   36.85   45.00   91.00     152
```

Graphs

I will then group them into age ranges to get a better sense of its distribution. As shown in the barchart below, the age range with the highest amount of police shootings is ages 25-35. The lowest is ages 85-95.

```r
police$age_range <- cut_width(police$age, width = 10)

ggplot(police, aes(x = age_range)) + geom_bar() + ggtitle("Distribution of police involved killings bas
    ylab("Number of police involved killings") + xlab("Age range")
```



Variable: Signs of Mental Illness

Summary Statistics

The table below compares the data for the categorical nominal variable, signs of mental illness. If they did not display signs of mental illness, it will be false. If they displayed signs of mental illness, it will be true.

```
table(police$signs_of_mental_illness)
```

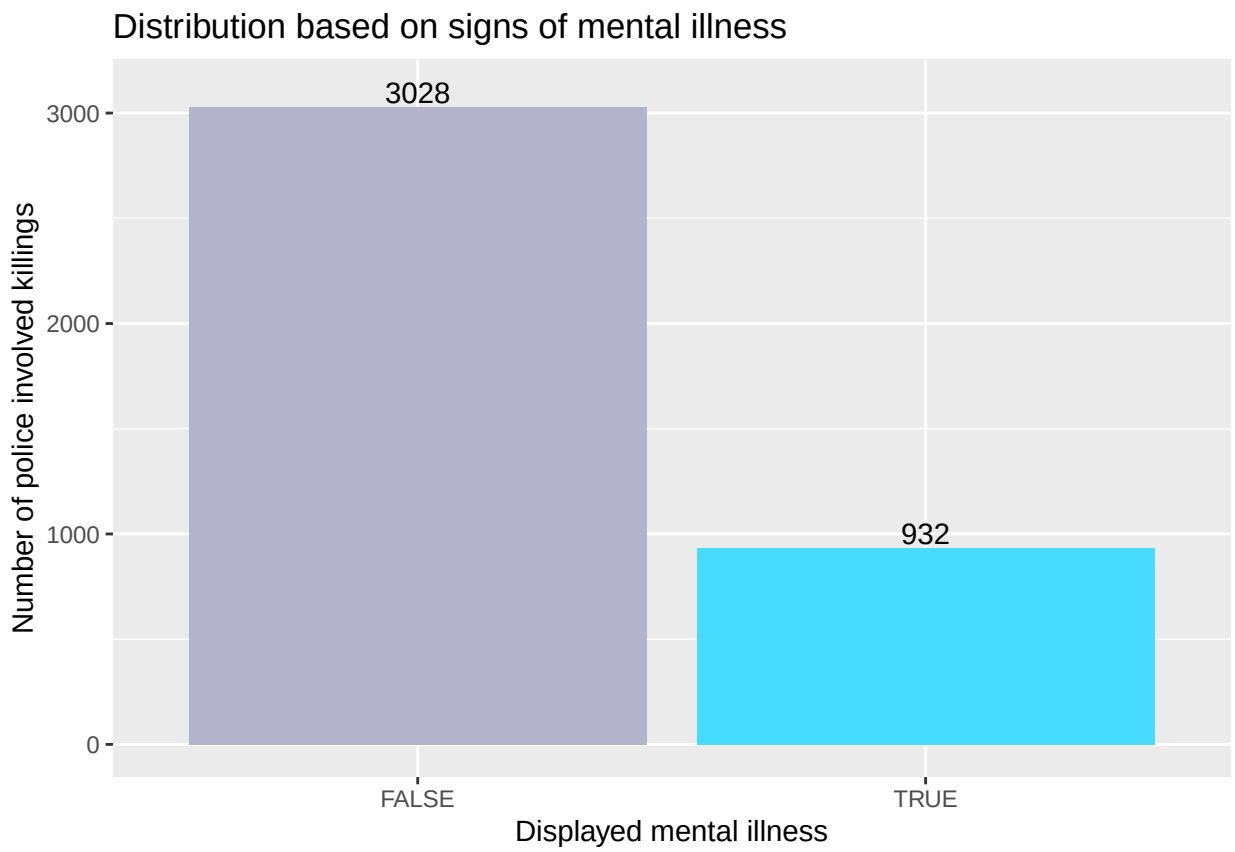
```
##
## FALSE  TRUE
##  3028   932
```

Graphs

As shown in the graph, there are more people who did not display signs of mental illness than those who did.

```
ggplot(police, aes(x = signs_of_mental_illness, fill = signs_of_mental_illness)) +
    geom_bar() + scale_fill_manual(guide = "none", values = c("#b3b3cc",
    "#47dbff")) + ggtitle("Distribution based on signs of mental illness") +
    ylab("Number of police involved killings") + xlab("Displayed mental illness") +
    geom_text(aes(y = ..count.. + 70, label = ..count..), stat = "count",
        size = 4)
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Bivariate Exploration

The table displays the column proportions of signs of mental illness by age ranges. According to the graph, the highest proportion of people who displayed signs of mental illness based on age range is 29.8% for ages 25-35. The lowest is 0.2% for ages 5-15 and 85-95. When calculating row proportions, however, we can see that in the category for displaying mental illness, there is an increasing trend of killings as age increases.

```
table(police$age_range, police$signs_of_mental_illness) %>%
    prop.table(margin = 2) %>%
    round(3)
```

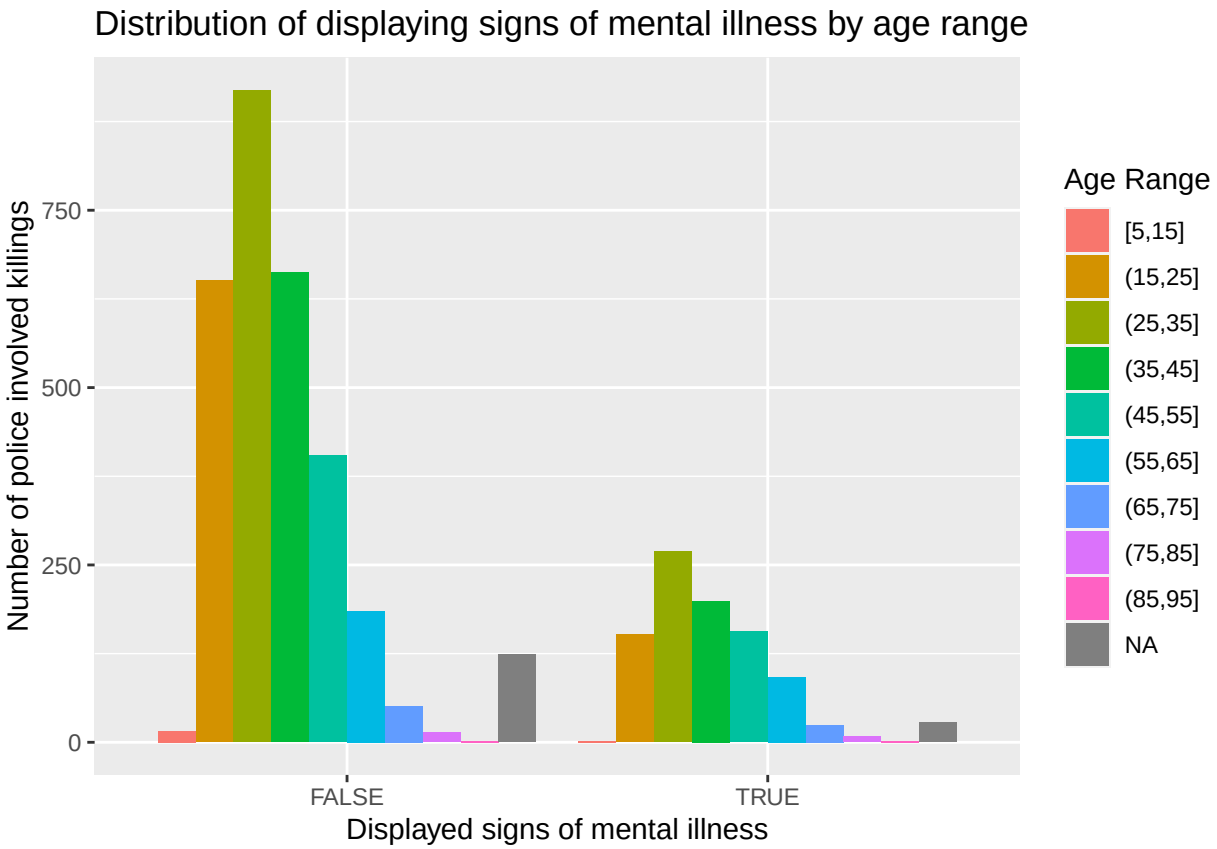
```
##
##           FALSE  TRUE
##   [5,15]  0.006 0.002
##   (15,25] 0.224 0.168
##   (25,35] 0.316 0.298
##   (35,45] 0.228 0.220
##   (45,55] 0.139 0.173
##   (55,65] 0.064 0.102
##   (65,75] 0.018 0.027
##   (75,85] 0.005 0.009
##   (85,95] 0.001 0.002
```

```
table(police$age_range, police$signs_of_mental_illness) %>%
    prop.table(margin = 1) %>%
    round(3)
```

```
##
##           FALSE  TRUE
##   [5,15]  0.889 0.111
##   (15,25] 0.811 0.189
##   (25,35] 0.774 0.226
##   (35,45] 0.769 0.231
##   (45,55] 0.721 0.279
##   (55,65] 0.668 0.332
##   (65,75] 0.680 0.320
##   (75,85] 0.636 0.364
##   (85,95] 0.500 0.500
```

The graph below is a barchart comparing the variables, signs of mental illness and age range. The graph is skewed right, showing that, on average, younger people tend to be killed by the police when displaying signs of mental illness than older people.

```
ggplot(police, aes(x = signs_of_mental_illness, fill = age_range)) +
    geom_bar(position = "dodge") + scale_fill_discrete(name = "Age Range") +
    ggtitle("Distribution of displaying signs of mental illness by age range") +
    ylab("Number of police involved killings") + xlab("Displayed signs of mental illness")
```



Conclusion

From my research, I found that my prior hypothesis is correct: Younger people tend to get killed more often by the police when displaying signs of mental illness than older people. However, we also found that when looking at the variable, age, itself, older people will tend to be killed than not if they display signs of mental illness. Younger people will tend to not be killed if they display signs of mental illness.