EDA_JulianneCaburian

Julianne Caburian

2024-09-23

1. Choose a dataset and read it into R.
2. Browse through the codebook

Introduction:

The data set I am using is washingtonpost/ data-police-shootings. Policeshootings is a database compiling every fatal shooting from 2015-2019 performed by an officer, by The Washington Post. I'll be exploring three variables in this Policeshootings database of how: 'age', 'race', and 'threat_level' which may relate to the outcomes of fatal encounters with law enforcement.

My research question in this project is "Does age, race, and threat level all have a connection?" In order to visualize and better understand police-shootings and its patterns. So is there an overall higher threat level with a specific age and race? I hypothesize that younger and older ages with all races will have less overall threat level.

Here I set up packages and set up a tibble up to 13 columns and 10 rows down so I can see up to the threat_level variable.For some reason I can't open up the dataset if I have it open on excel as well.

```
library(readxl)
fps  <- read_excel("C:/Users/cabur/OneDrive/Desktop/fatal-police-shootings-data.xlsx", sheet=1, col_nam
fps[1:10,1:12]
```

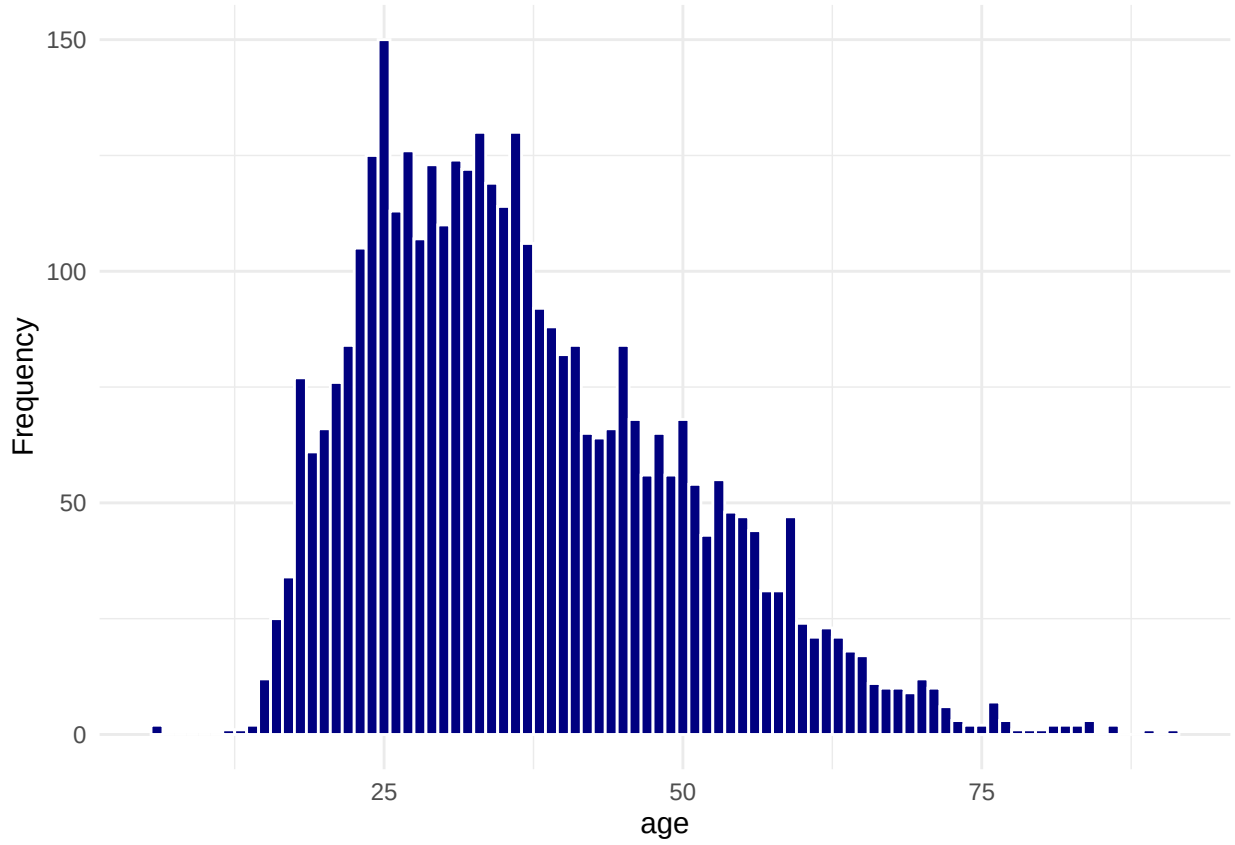
```
## # A tibble: 10 x 12
##       id name       date                manner_of_death armed   age gender race
##    <dbl> <chr>      <dttm>              <chr>           <chr> <dbl> <chr>  <chr>
##  1     3 Tim Elliot 2015-01-02 00:00:00 shot            gun      53 M      A
##  2     4 Lewis Lee~ 2015-01-02 00:00:00 shot            gun      47 M      W
##  3     5 John Paul~ 2015-01-03 00:00:00 shot and Taser~ unar~    23 M      H
##  4     8 Matthew H~ 2015-01-04 00:00:00 shot            toy ~    32 M      W
##  5     9 Michael R~ 2015-01-04 00:00:00 shot            nail~    39 M      H
##  6    11 Kenneth J~ 2015-01-04 00:00:00 shot            gun      18 M      W
##  7    13 Kenneth A~ 2015-01-05 00:00:00 shot            gun      22 M      H
##  8    15 Brock Nic~ 2015-01-06 00:00:00 shot            gun      35 M      W
##  9    16 Autumn St~ 2015-01-06 00:00:00 shot            unar~    34 F      W
## 10    17 Leslie Sa~ 2015-01-06 00:00:00 shot            toy ~    47 M      B
## # i 4 more variables: city <chr>, state <chr>, signs_of_mental_illness <lgl>,
## #   threat_level <chr>
```

Univariate Exploration: Describe each of the variables under consideration.

Here I'm going to do a Univariate Analysis with histograms and barplots to visualize the data from 'fps' (fatal police shootings. So this means calculating some summary statistics (N(%) or mean(sd)) and make a graphic (histogram and barchart)

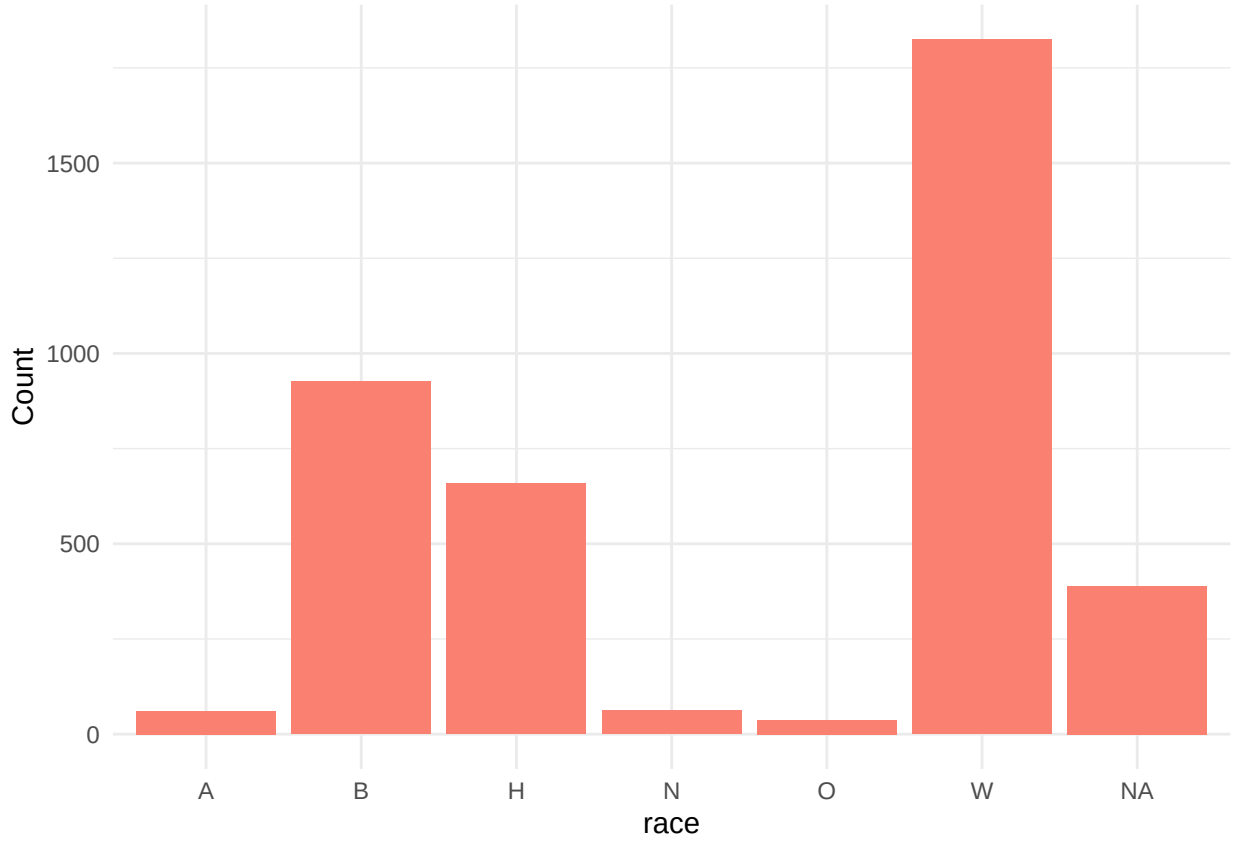
this is going to use ggplot2 and dplyr, each dataset has a summary/and or table for exact numbers outside of the visual.

```
## Warning: Removed 152 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```




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

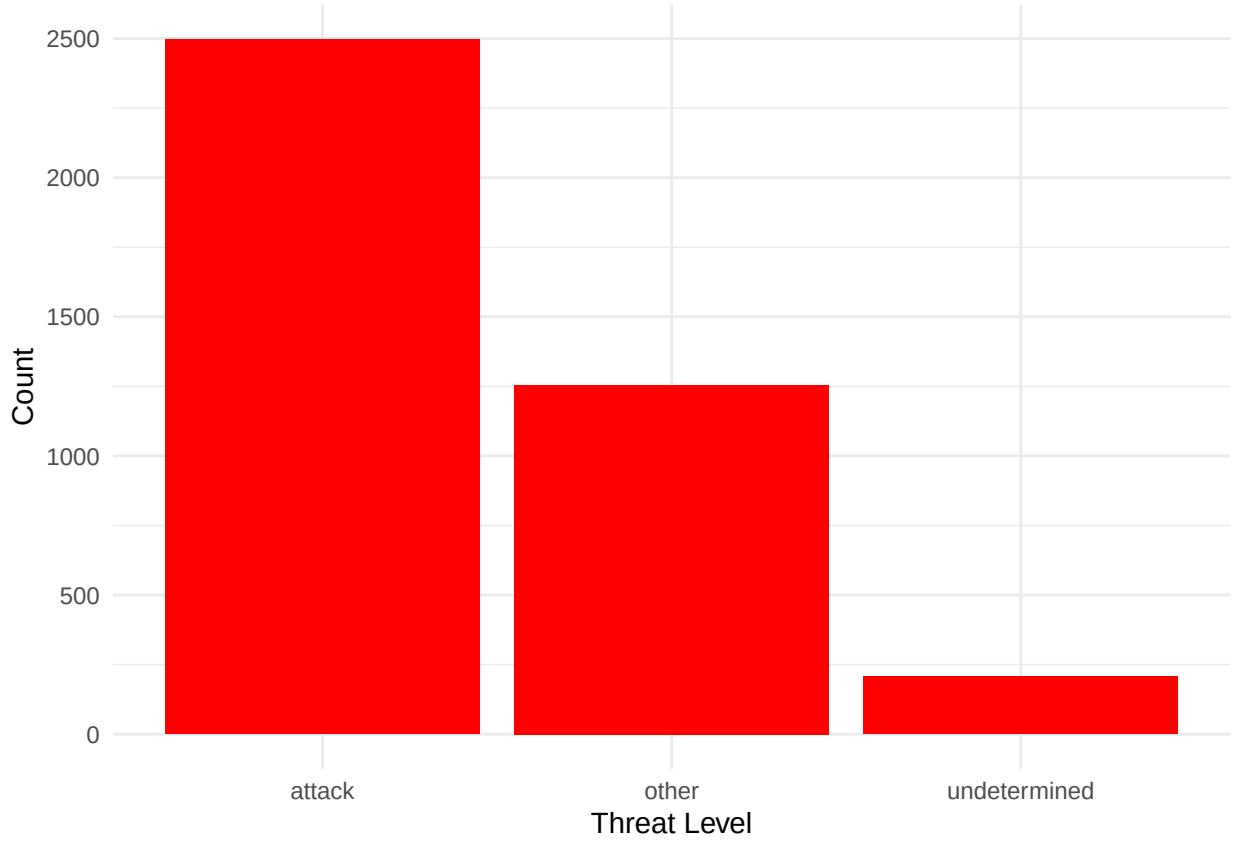
```
fps %>%
  ggplot(aes(x=race))+ geom_bar(fill="salmon") + xlab("race")+ ylab("Count") + theme_minimal()
```




```
table(fps$race)
```

```
##
##    A    B    H    N    O    W
##   61  927  659   62   37 1825
```

```
fps %>%
  ggplot(aes(x=threat_level))+ geom_bar(fill="red") + xlab("Threat Level")+ ylab("Count") + theme_minima
```




```
table(fps$threat_level)
```

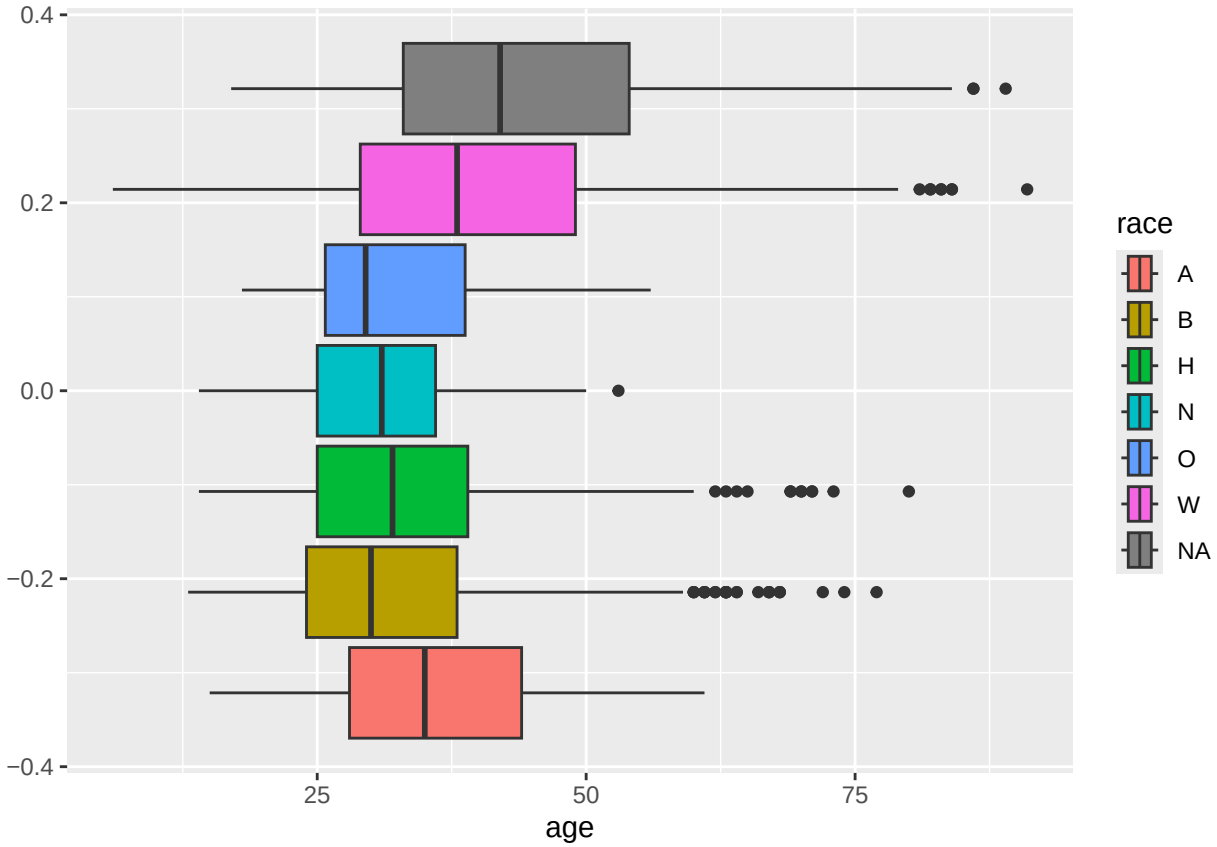
```
##
##       attack        other undetermined
##         2497         1255          208
```

Bivariate Exploration

3.

```
ggplot(fps, aes(x=age, fill = race)) + geom_boxplot()
```

```
## Warning: Removed 152 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

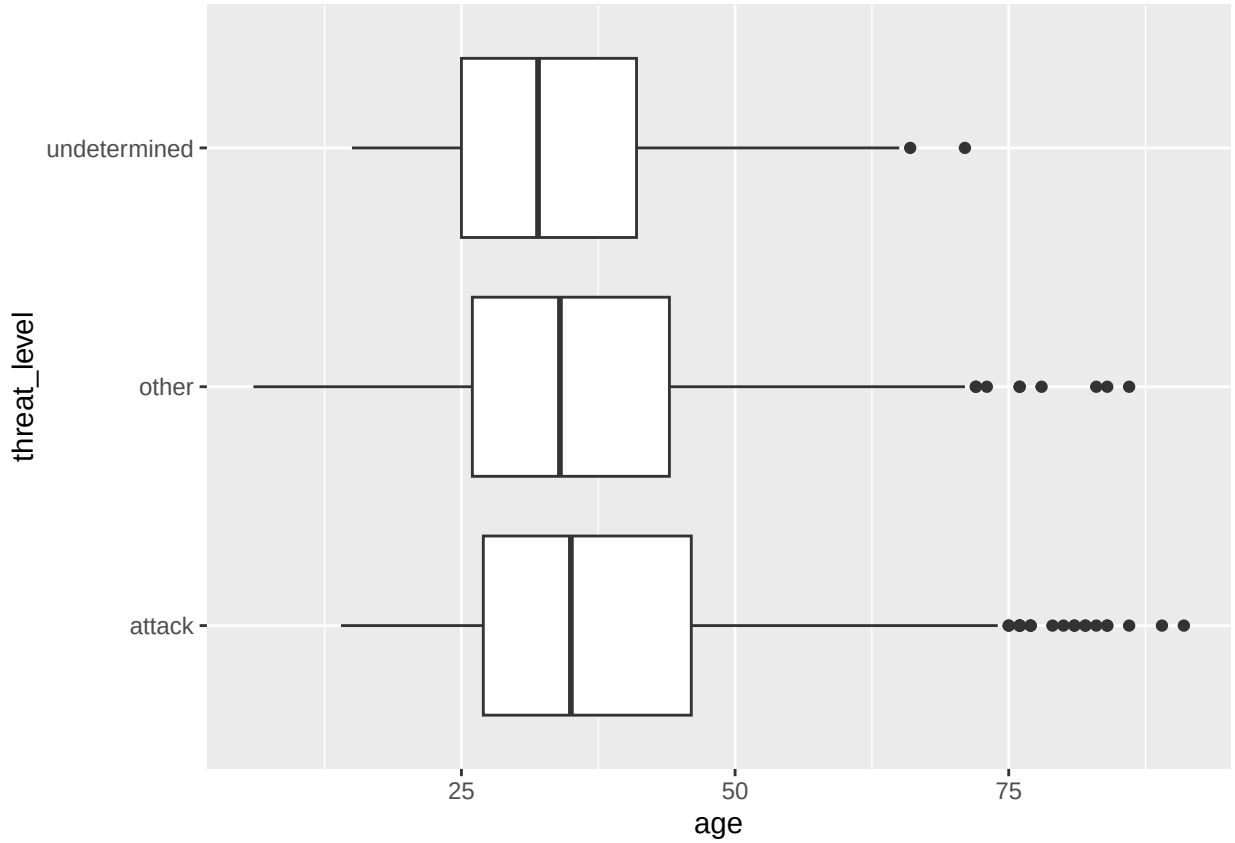


A (asian), B(black/african american), H(hispanic/latino), N(native american), O(other), and W(white)

This boxplot shows the age of each race that was shot, we cab see that for when no race was identified, the median was higher, but those whom were white were on average older, while the youngest ones were black.

```
ggplot(fps, aes(y=threat_level, x=age)) + geom_boxplot()
```

```
## Warning: Removed 152 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



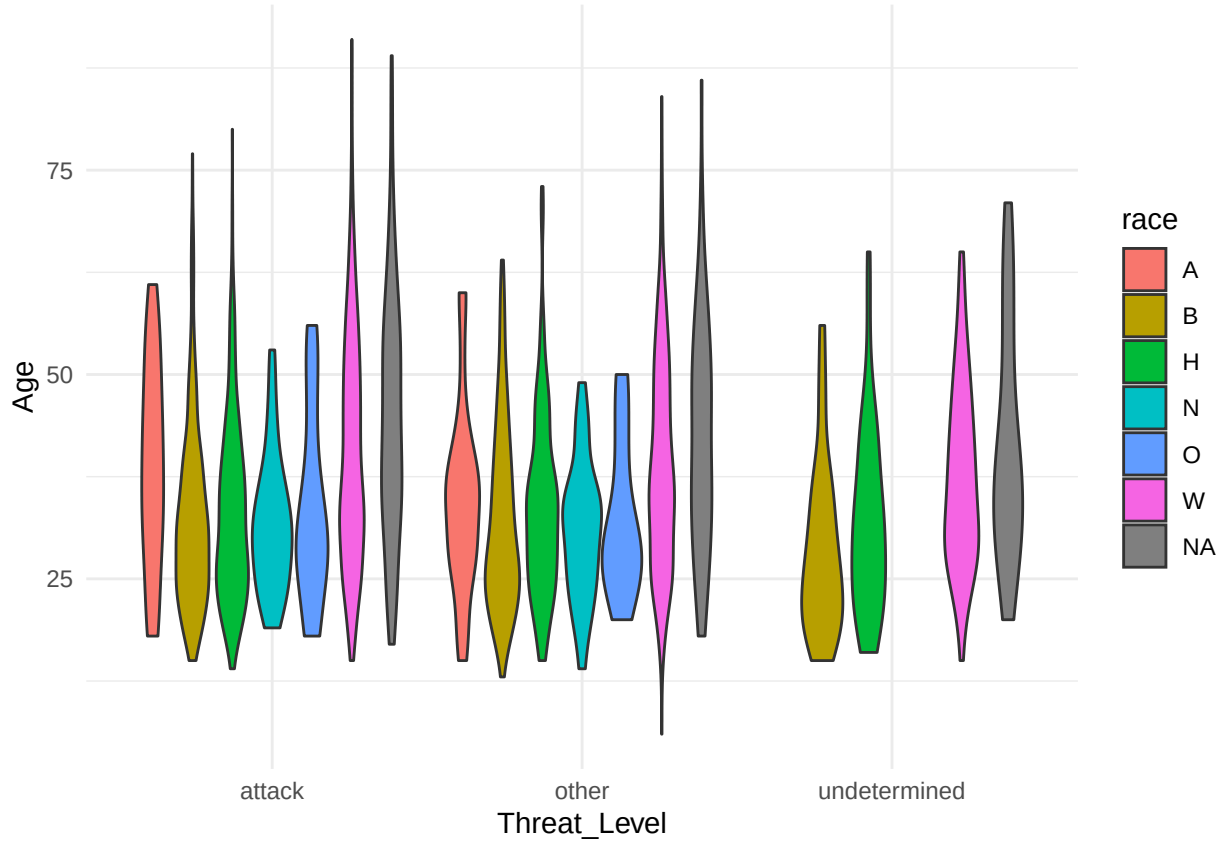
For the threat level to age comparision boxplot... the IQR's on the interquartile range have a Q1 that roughly starts at early 20's, and Q3 right around the 30-40 range. The median range is all roughly the same for each age as well, 33-37 years old.

Here im going to use a violin plot to see all three variables together.

```
ggplot(fps, aes(x=factor(threat_level), y = age, fill = race)) +
  geom_violin(drop = FALSE) +
  xlab("Threat_Level") +
  ylab("Age") +
  theme_minimal()
```

```
## Warning: Removed 152 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Cannot compute density for groups with fewer than two datapoints.
```




```
table(fps$threat_level, fps$race) %>%
  prop.table
```

```
##
##                            A           B           H           N           O
##   attack       0.0089610753 0.1705404649 0.1033323999 0.0098011761 0.0067208065
##   other        0.0075609073 0.0758891067 0.0691683002 0.0072808737 0.0036404369
##   undetermined 0.0005600672 0.0131615794 0.0120414450 0.0002800336 0.0000000000
##
##                            W
##   attack       0.3326799216
##   other        0.1568188183
##   undetermined 0.0215625875
```

The violin plot shows that there's a lower threat level at younger ages, and same for the old ones. There's also longer violins where you can see that age is variable for a race within it's own threat level. We can see this in the longer ones.

The table lists percentages so we can see the rates of each race and their threat_level data. W, (white) shows up to have 33% of threat_level of attack while the least is in O (other), which is 0.6%.

Conclusion My original hypothesis was that younger and older ages with all races will have less overall threat level. This is true! I used a violin plot in order to show all three so we could visualize all three variables together. Those from ages 27-45 throughout each race are more likely to have a higher threat_level, and higher rate of being in an accident. It varies for specific races as well, as we can see that in the age and race boxplot, white people had the highest median of incidents. So even though my hypothesis was true for

each race, some races have smaller and larger ranges. Regarding my research question, "Does age, race, and threat level all have a connection?" Yes. All races will have that range around 27-45, and threat level is generally higher for the white race.