Exploratory Data Analysis

Aaron David

2/22/2021

Introduction - Police Shootings

The data set I will be analyzing is the Police Shooting data set. This data set is a study of fatal police shootings in the United States. The data was collected through the Washington Post and is a record of all shootings since January 1st of 2015. The data only includes reports of on-duty officers in which the officers killed the individuals by shooting. In this Exploratory Data Analysis I will analyze the following variables: the threat level of the individual killed, the ages of the individuals, and the races of the individuals.

```
police <- read_excel("/math130/data/fatal-police-shootings-data.xlsx", sheet=1,
                     col_names=TRUE)
```

Univariate Analysis Of Variables

Threat Level of Victim

```
police$threat_level_new <- fct_recode(police$threat_level, "Attack Threat" =
                                      "attack", "Significant Threat" = "other",
                                      "Undetermined" = "undetermined" )

table(police$threat_level_new)
```

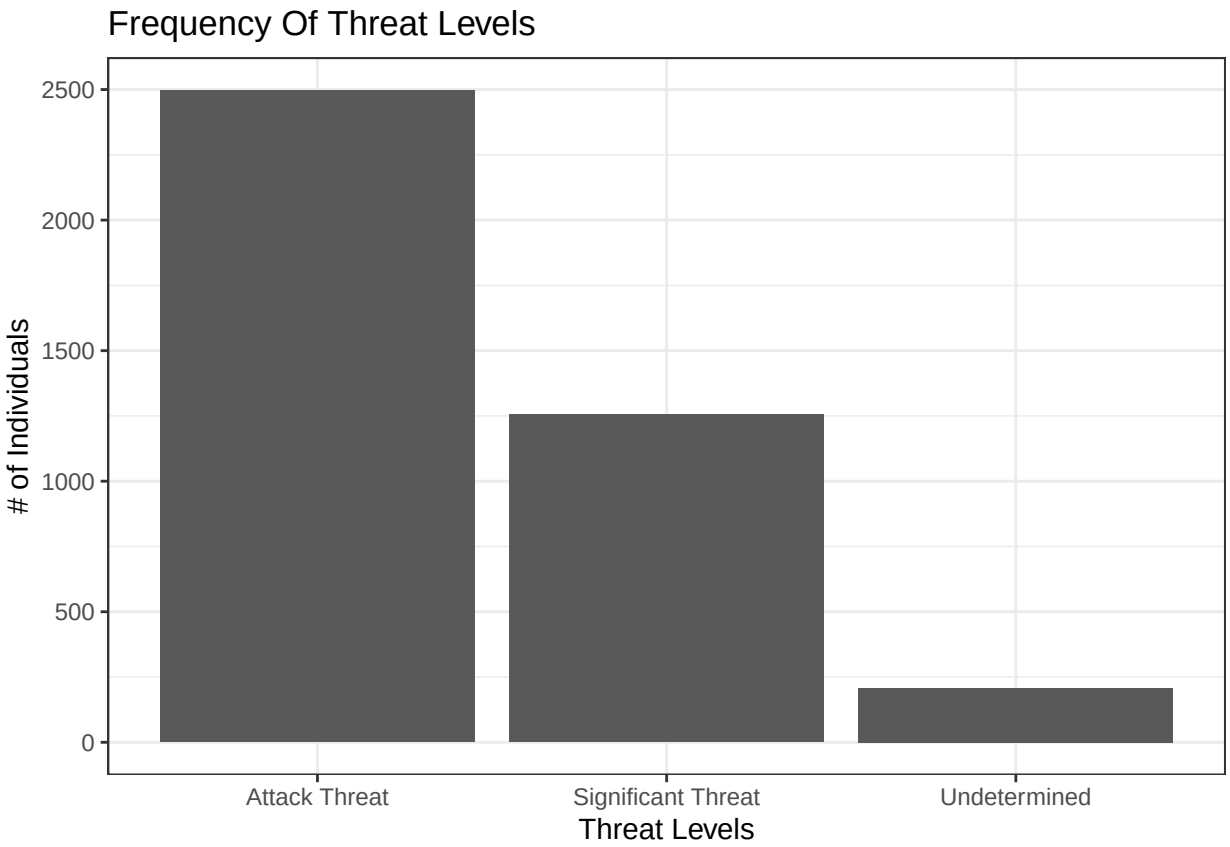
```
##
##      Attack Threat Significant Threat       Undetermined
##               2497               1255                208
```

I modified the names of the different threat levels as I felt 'other' was too vague and did not make sense given that it meant the individual posed a significant threat.

In this table "attack" refers to if the individual was physically attacking the officer, "significant threat" refers to if the individual simply posed a significant threat to the officer, and "undetermined" means that the threat of the individual could not be determined.

As the above table shows, 2497 of the victims were physically attacking the police officers prior to their deaths, 1255 of the victims posed a significant threat to the officers, and 208 of the victims threat level could not be determined.

```
ggplot(police, aes(x=threat_level_new)) + geom_bar() + theme_bw() +
  xlab("Threat Levels") + ylab("# of Individuals") +
  ggtitle("Frequency Of Threat Levels")
```



This sizes of the bars show the imbalance of individuals determined to be attacking the officers rather than those simply posing a significant threat to the officers. The height of the 'attack' threat bar is about twice as tall as the height of those who posed a significant threat.

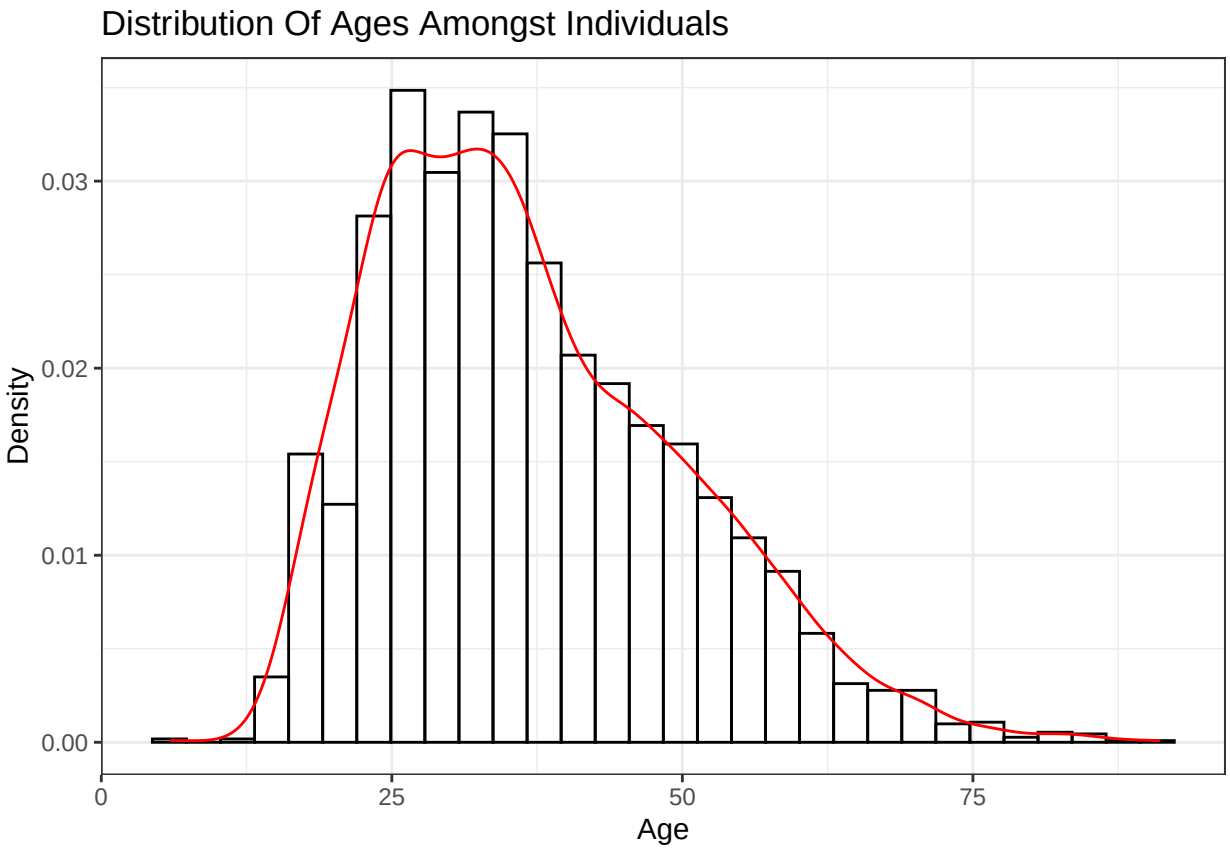
Age of Individuals

```
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
```

This summary statistics shows that the mean age of individuals killed was about 37. The youngest individual shot and killed was 6 years of age and the oldest was a 91 year old. It should be noted that there are 152 NA values which likely suggests that the ages of those individuals was never released.

```
ggplot(police, aes(x=age)) + geom_histogram(aes(y=..density..),
                                            col = 'black', fill=NA) +
  geom_density(col='red') + theme_bw() + xlab("Age") + ylab("Density") +
  ggtitle("Distribution Of Ages Amongst Individuals")
```



This histogram depicts the distribution of all individuals ages, except of course the NA values, with an overlaid density plot to help show the shape of the distribution. According to this graph, the majority of individuals have an age approximately between 25 and 50 years.

Race

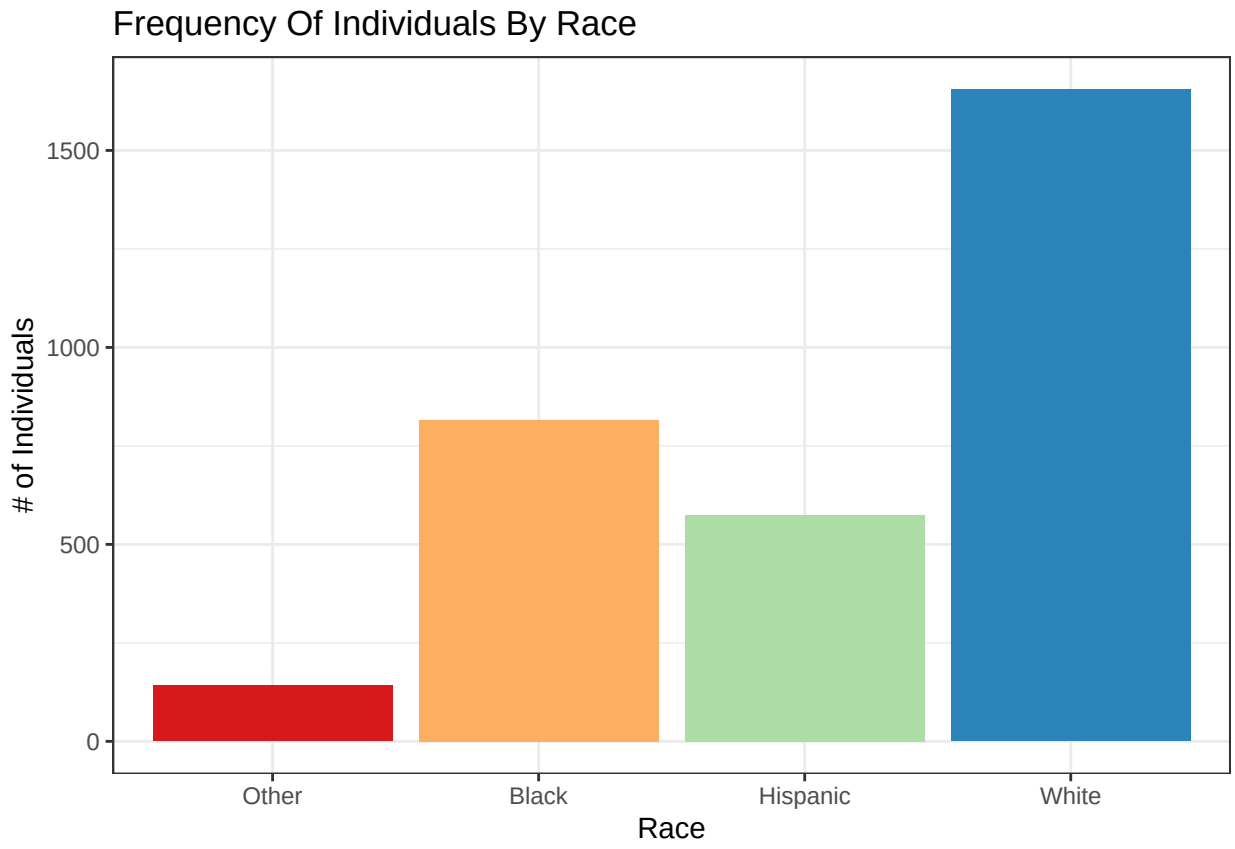
```
police$race <- fct_recode(police$race, "Asian" = "A", "Black" = "B",
                          "Hispanic" = "H", "Native American" = "N",
                          "Other" = "O", "White" = "W")
table(police$race)
```

```
##
##            Asian            Black         Hispanic Native American
##               61              927              659               62
##            Other            White
##               37             1825
```

I renamed the factors to clearly see the races of the individuals. From this table we can see that a majority of the individuals were White, Black, and Hispanic. There are about twice as many White individuals killed as Black individuals and about a third as many Hispanic individuals as White. I'm not too interested in the Asian and Native American races since they make up such a small portion of the data, so I will lump them into 'other' to shift focus to White, Hispanic, and Black races.

```
police$race_new <- fct_collapse(police$race, "Other" =
                                  c("Asian", "Native American", "Other"))

police %>% na.omit() %>% ggplot(aes(x=race_new, fill=race_new)) +
  geom_bar() + theme_bw() + xlab("Race") + ylab("# of Individuals") +
  ggtitle("Frequency Of Individuals By Race") +
  scale_fill_brewer(palette="Spectral", guide=FALSE)
```

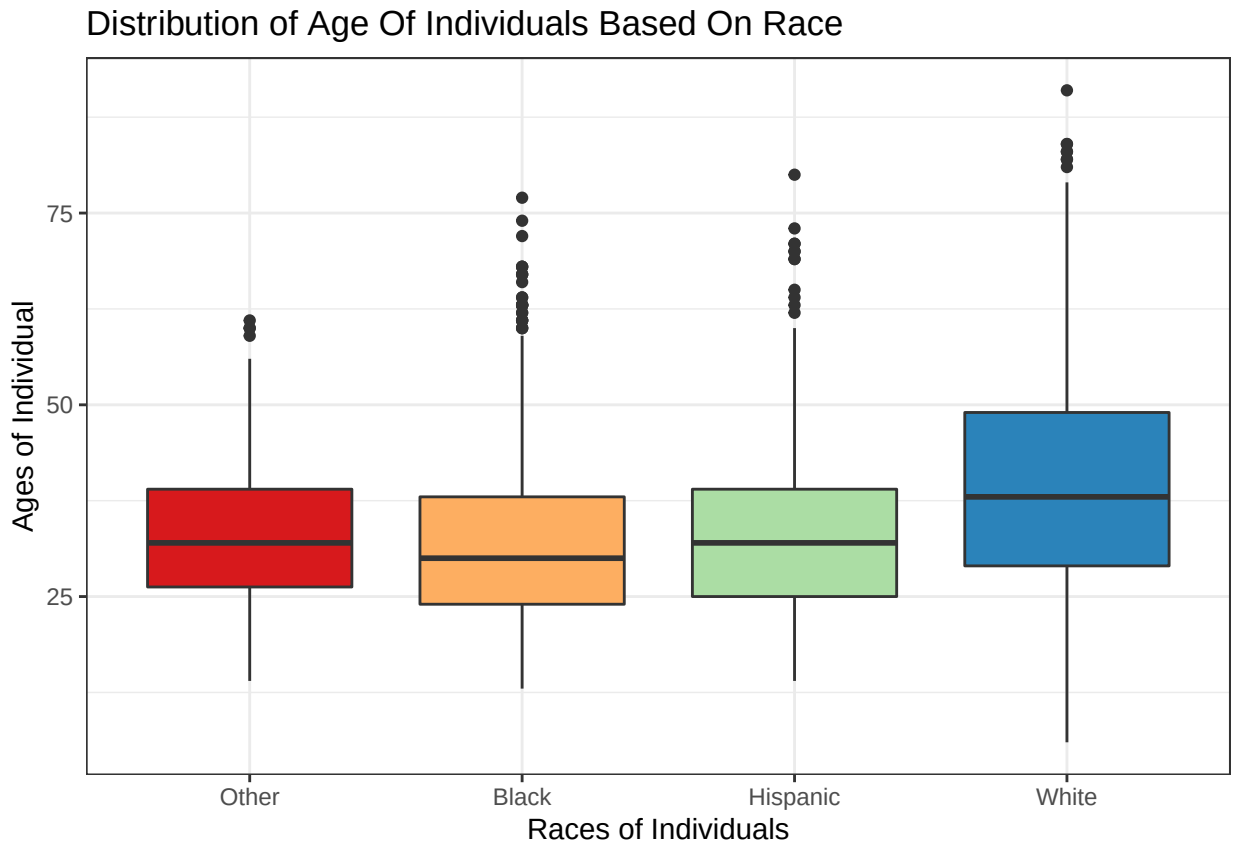


This bar chart shows the frequency of individuals shot by race. I added a fill to help distinguish and make it slightly easier to compare the frequency of each race. As expected by the table I made beforehand, the height of the bar representing White individuals is about twice as high as the bar of the Black individuals.

Bivariate Analysis

Age & Race

```
police %>% filter(race_new=="Black"|race_new=="Hispanic"|
                    race_new=="White" | race_new=="Other") %>%
  ggplot(aes(y=age, x=race_new, fill=race_new)) + geom_boxplot() + theme_bw() +
  xlab("Races of Individuals") + ylab("Ages of Individual") +
  ggtitle("Distribution of Age Of Individuals Based On Race") +
  scale_fill_brewer(palette="Spectral", guide=FALSE)
```



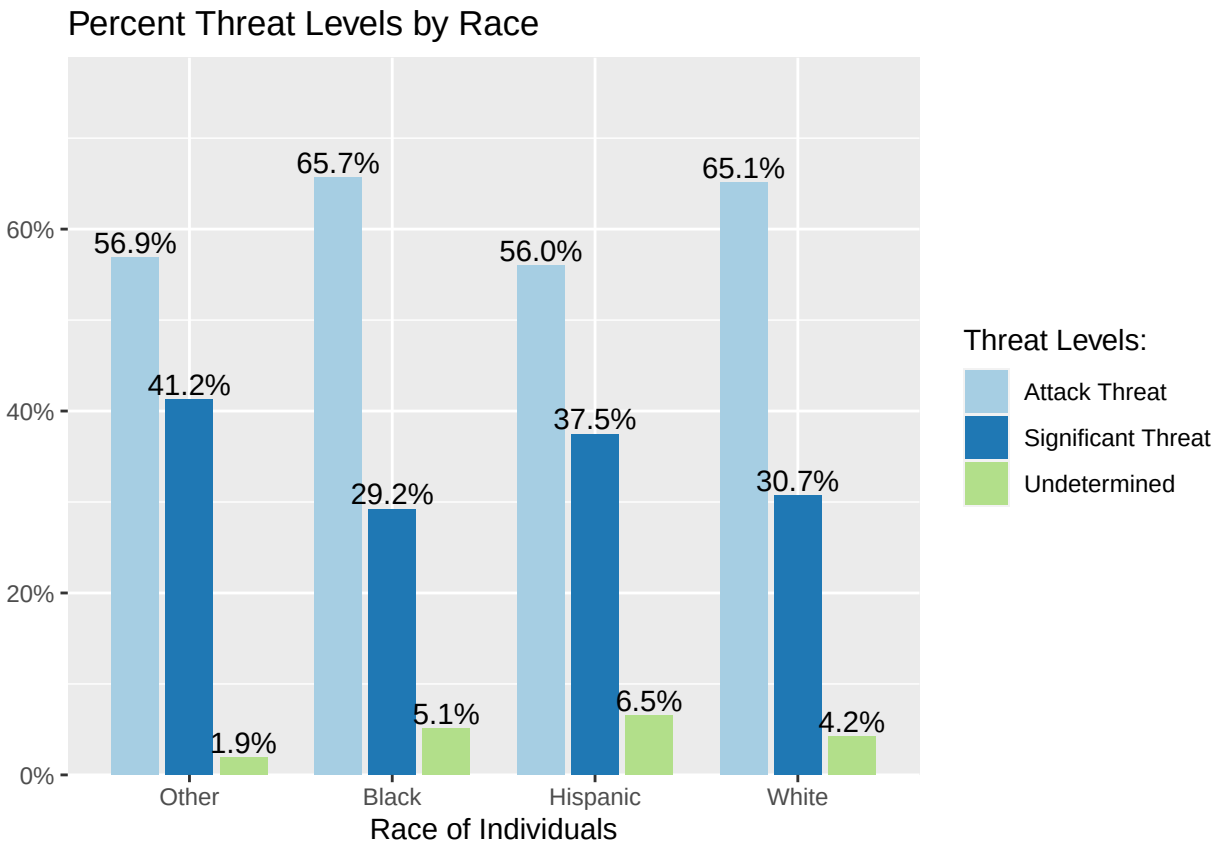
According to this graph there is little to no correlation between the race of the individuals and the age at which they were shot. White individuals on average tend to be marginally older than all other races however. The whiskers of the box plot representing white individuals show that the oldest and youngest individuals shot were white.

Race & Threat Level

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

```
##
##            Attack Threat Significant Threat Undetermined
##   Other             0.569              0.412        0.019
##   Black             0.657              0.292        0.051
##   Hispanic          0.560              0.375        0.065
##   White             0.651              0.307        0.042
```

```
plot_xtab(police$race_new,police$threat_level_new, margin='row', show.total=F,
          legend.title = "Threat Levels:", show.n=F, title=
            "Percent Threat Levels by Race") +
  xlab("Race of Individuals")
```



This graph helps visualize the proportion of each threat level within race. The graph helps point out some trends that were at first not very apparent in the prop. table. Surprisingly, despite Hispanic individuals being less likely to be labeled as an attack threat, they are about 7% more likely than White and Black individuals to be shot while only posing a significant threat to the officers. Also surprising is that Black individuals and White individuals are almost equally likely to be shot and killed regardless of their individual threat level.