The Disproportions Within Fatal Police Shootings

Jerome Veix, Paula Palles, Williem Christian Encarnacion, and Zacharia Mwaura

Minor in Data Science

219:220 Fundamentals of Data Visualization in R: Final project

Prof Bruno Richard, PhD

Introduction to Dataset:

The Police Shootings dataset shows the reports of fatalities by police departments, newspapers, and social media as collected by the Washington Post since Jan 1, 2015. The media phenomenon was inspired to collect this data following the murder of a defenseless, unarmed black man, known as Freddie Gray, in 2014 in Ferguson, Missouri. His unfortunate death symbolized yet another fatal interaction between a black person and the institutional powerhouse of the police force. Within the data frame of victims, seventeen different columns consist of geographical, medical conditions, social markers of victims, and more.

Our Research Questions:

1. What racial groups are mostly represented by the victims in the dataset?
2. How can different social markers impact a victim of police brutality being labeled with a negative connotation such as an attack threat level?
3. Is there any correlation between an attack threat level and an increased rate of death from law enforcement?

In order to begin our analysis, we needed to create our Hypotheses.

Generating Our Hypothesis:

Null Hypothesis: There is no difference between race & gender and other columns on their impact in skewing the attack threat level for a group.

Alternate Hypothesis: Black males will have a higher probability of being described with an attack threat level than their white counterparts.

After we created our hypotheses we could now begin our exploratory data analysis (EDA).

Beginning of EDA - Data Cleaning:

The raw dataset had 6953 rows but after omitting the NA values, we worked with 6121 rows. The data sets also consisted of blank values which we left in analyzing the data.

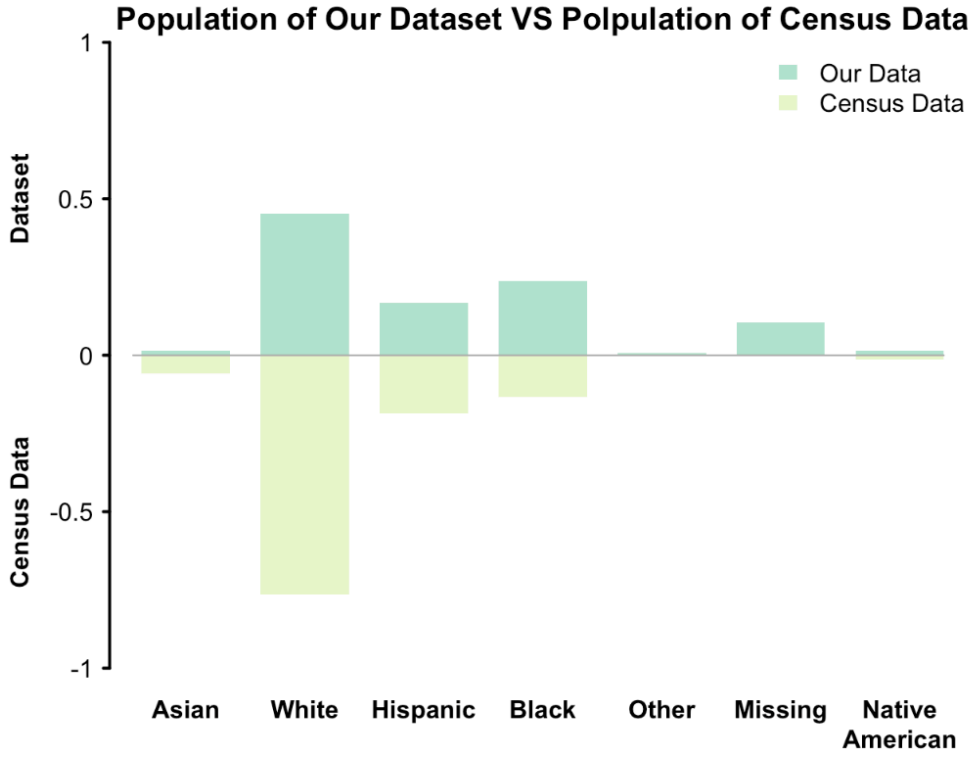
	id <int>	name <chr>	date <chr>	manner_of_death <chr>	armed <chr>	age <int>	gender <chr>	race <chr>	city <chr>
1	3	Tim Elliot	2015-01-02	shot	gun	53	M	A	Shelton
2	4	Lewis Lee Lembke	2015-01-02	shot	gun	47	M	W	Aloha
3	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23	M	H	Wichita
4	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32	M	W	San Francisco
5	9	Michael Rodriguez	2015-01-04	shot	nail gun	39	M	H	Evans
6	11	Kenneth Joe Brown	2015-01-04	shot	gun	18	M	W	Guthrie
7	13	Kenneth Arnold Buck	2015-01-05	shot	gun	22	M	H	Chandler
8	15	Brock Nichols	2015-01-06	shot	gun	35	M	W	Assaria
9	16	Autumn Steele	2015-01-06	shot	unarmed	34	F	W	Burlington
10	17	Leslie Sapp III	2015-01-06	shot	toy weapon	47	M	B	Knoxville

1-10 of 6,121 rows | 1-10 of 17 columns

Previous 1 2 3 4 5 6 ... 100 Next

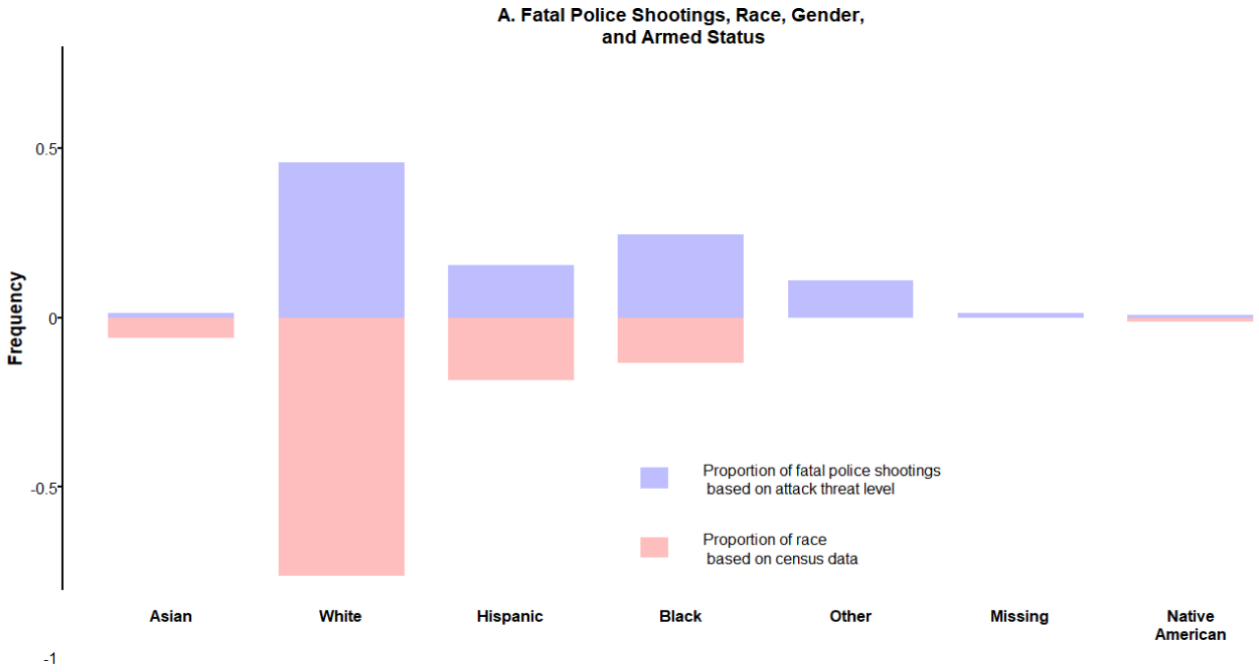
Key Finding → Percentages of Racial Markers Vs Federal Census Data

To establish the populations of each group we created a for loop. With this for loop, we created a graph that combines the dataset's reports on each racial group's presence and the federal census on the country's racial demographics. It showed us that black people are over-represented in this data set and have a higher percentage of fatalities. They make up a higher population in the dataset than in real life. The dataset doesn't take into account that there are higher populations of white people so that's why there are more discrepancies than what meets news coverage.



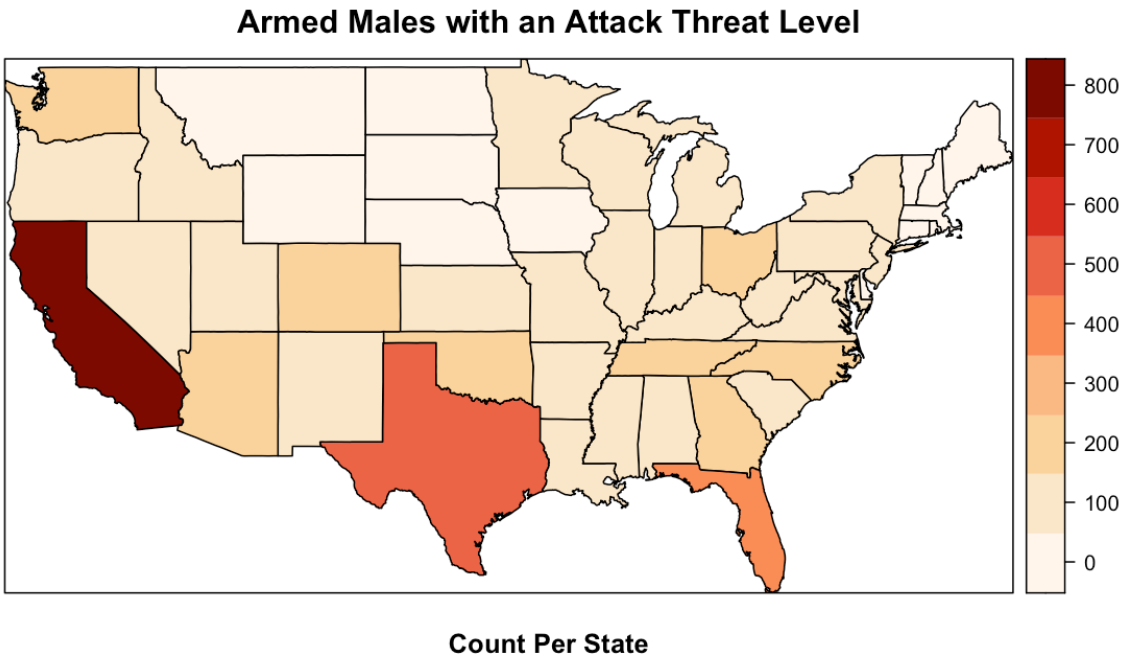
To dig deeper into this, we tried to add more factors from the dataset to see if they would affect the graph.

In the graph below we added gender and armed status to race.



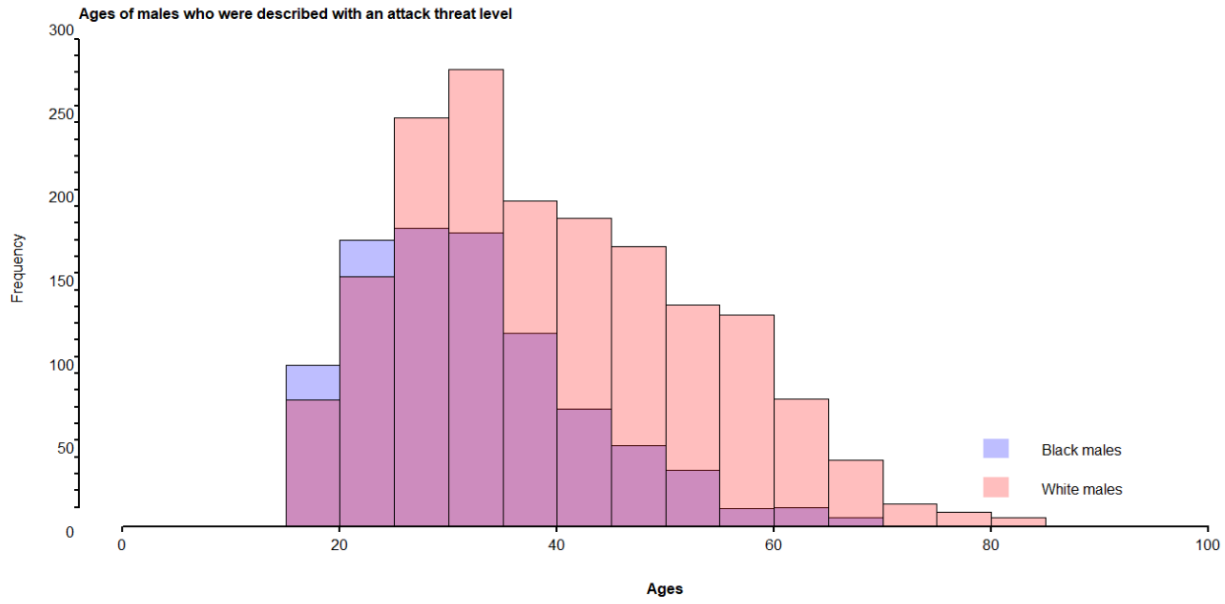
After we saw the graph with added factors it showed essentially the same result – same trends

To further our EDA we created a heat map:



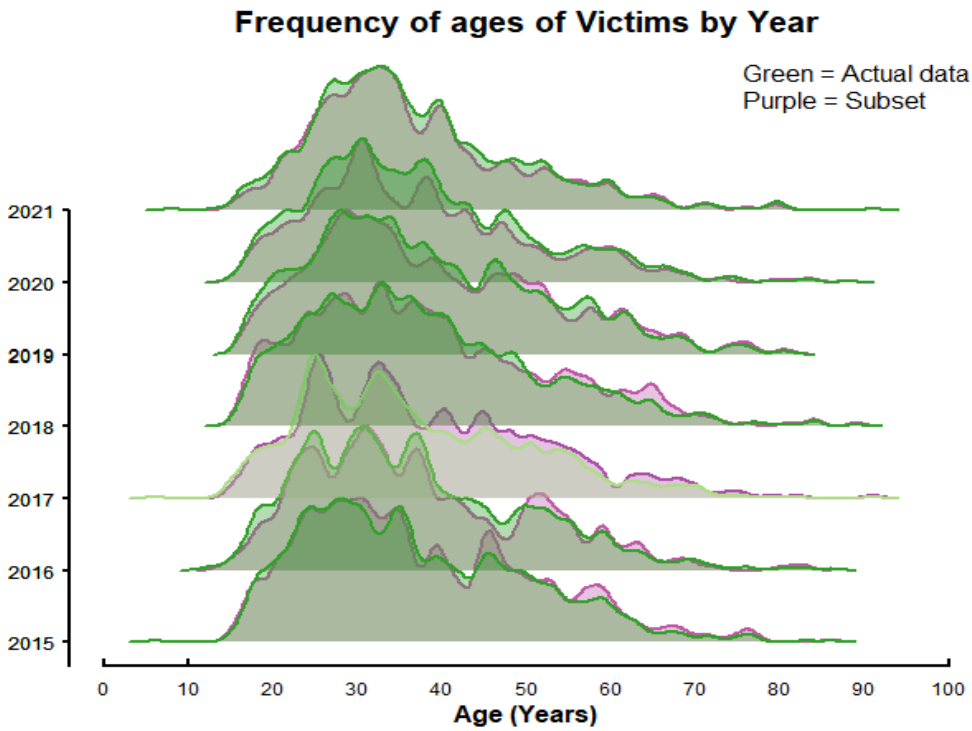
In this graph, we filtered the dataset for an attack threat level and counted each of the fatalities, per state, in the continental United States.

Another graph we created as a result of our EDA was creating a histogram:



In order to make this graph, we first filtered the dataset for race, specifically black and white, and for gender, male. With this, we took all the new values and created a histogram based on the distribution from the age column.

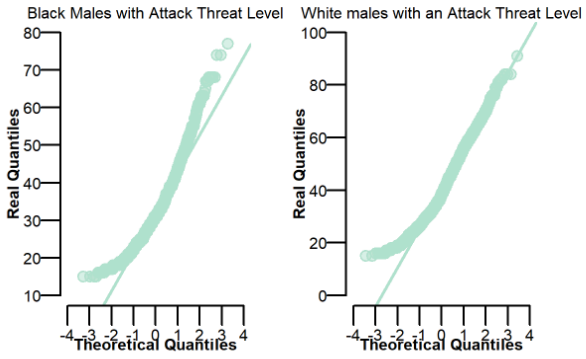
Concluding Out Our EDA we created another graph factoring for age with a joyplot:



Verifying the age factor's credibility in deep statistics:

Then, we verified the age factor's credibility in deep statistics through QQ plots, Levene's test, and the Welch Test. The QQ plots show how for the most part, the distribution of ages for both black males and white males with an attack threat level was normal. However, there were more points of departure for black people at the tails. Initially, we ran a t-test, but Levene's Test contained significant codes meaning we should do a Welch test instead. Also, this makes sense because the Welch test is for testing variables that have different variances in the two groups. The Welch test supports the findings of the histogram as the mean of y, which characterizes the ages of white males with an attack threat level, is higher than the mean of x, the black counterparts. The average for whites was 40.40, while for blacks, it was 32.81. It is good that the two different groups have different values and elements, but we must also contextualize the data found with accurate racial proportions.

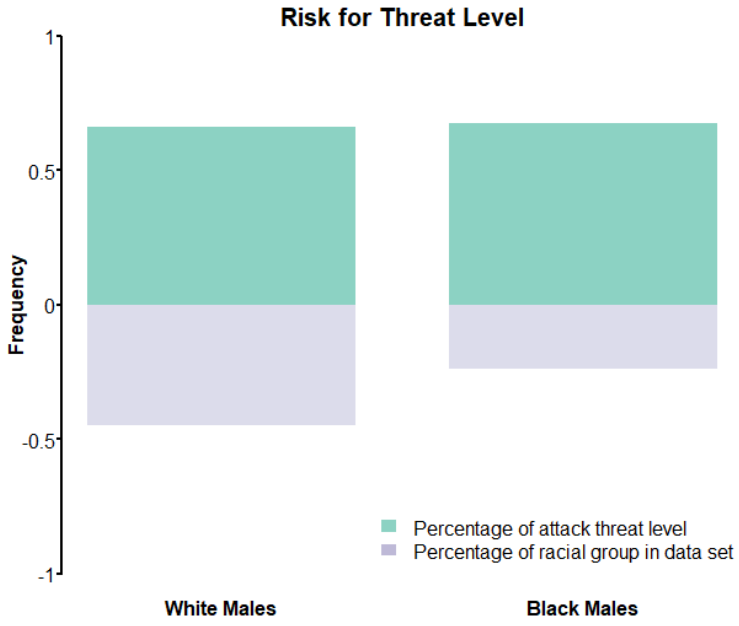
```
        Welch Two Sample t-test

data:  blackfolk$age and whitefolk$age
t = -15.579, df = 2340.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -6.788662
sample estimates:
mean of x mean of y
 32.81646  40.40683
```



```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1    87.9 < 2.2e-16 ***
##       2674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

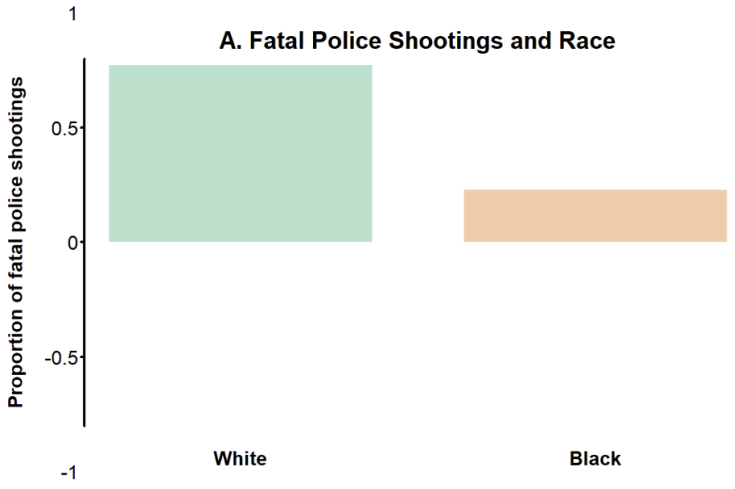
Answering our Alternative Hypothesis:

In the following bar graph, we outright answered our alternate hypothesis by proving how black males have a higher probability of achieving an attack threat level than their white counterparts. The racial groups described with an attack threat level should be divided by the group numbers to calculate the rate per race group. Initially, we thought we should compare an individual with their group because it would have more societal impact/accuracy. We concluded that black males slightly had a higher percentage of an attack threat level with 67% than white males, which garnered a 66%. On the surface, there isn't a substantial enough difference between these percentages to distinctly prove our hypothesis, hence the need to factor in the percentage of racial groups in the data set. The percentages for the victim are relative to white males as they comprised about 45% of the dataset with the 66% rate within the victim list of an attack threat level. In juxtaposition, black males only make up 24% of the data set but garnered an astounding 67% rate attack threat level which exceeds their expected proportional likelihood.



Modeling and Analysis:

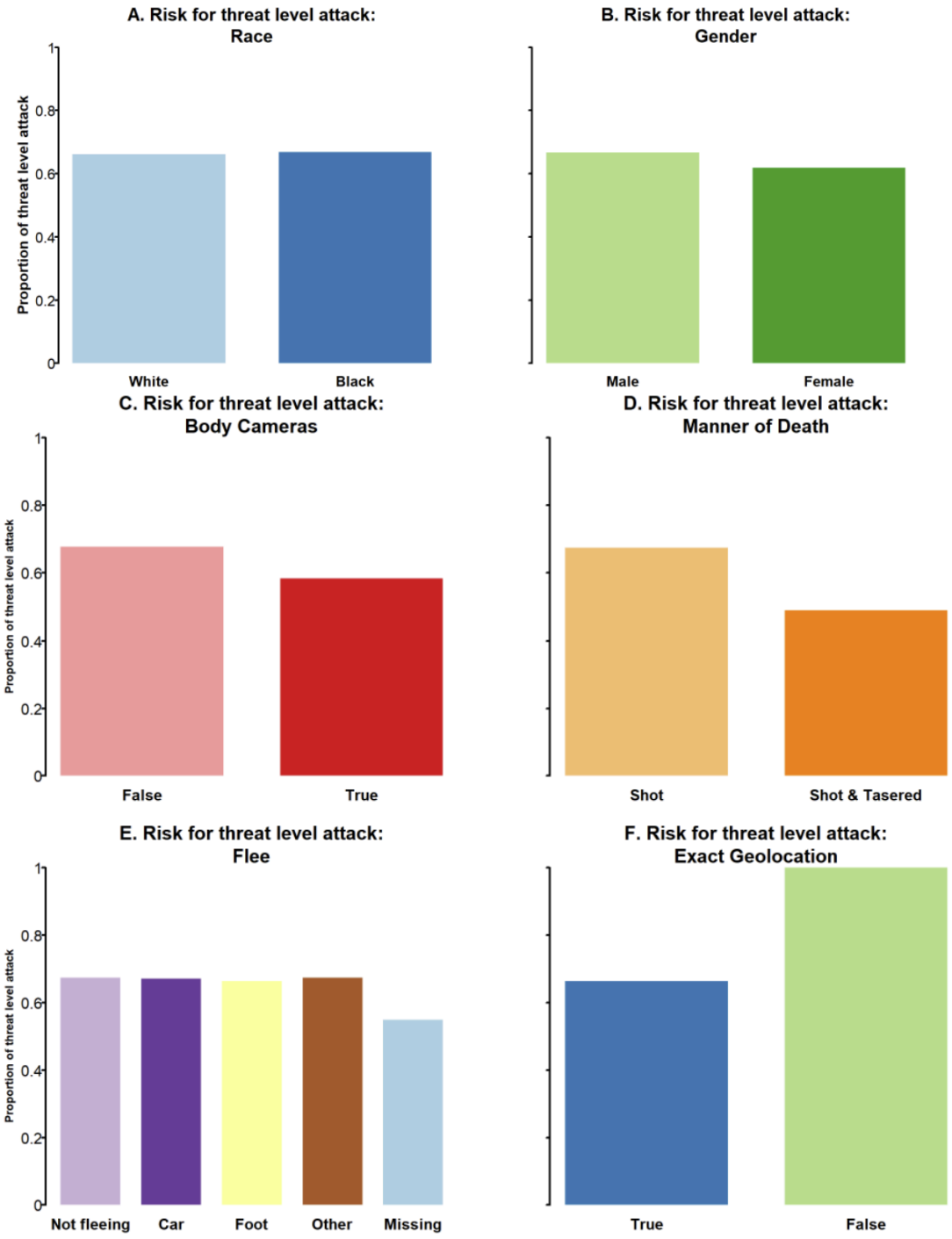
We tried to do the stratified random sampling but it just didn't work for us and we didn't know how to model it since it only contained two columns, value, and race. However, with a bit more time and less stress for finals, we definitely could've made it work and fixed it because we just had to figure out how to include the rest of the data in there.



Initially, in our PowerPoint presentation, we used threat level as our outcome variable. However, as we did the assignment and got feedback, we started to think the threat level could help determine who was black and who is white to certainly find discrepancies in fatal police shootings. However, we did it the other way around. In regards to our original hypothesis, we created 5 models after splitting our data into testing and training data to see how well we could predict if someone was a threat or not. Below are the results:

```
Call:
glm(formula = TLnumerical ~ gender + race, family = "binomial",
    data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5074  -1.4870   0.8800   0.8966   0.9935

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.44934    0.15960   2.815  0.00487 **
genderM      0.25433    0.16341   1.556  0.11962
race1        0.04524    0.08013   0.565  0.57233
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3972.9  on 3129  degrees of freedom
Residual deviance: 3970.1  on 3127  degrees of freedom
AIC: 3976.1
```

As you can see, based on our hypothesis, we have a very weak model, it's very liberal and the model is basically just reading the scores off the true answers and basically isn't doing anything. There also aren't many statistically significant factors as seen by the p-values of the summary of the model. The same goes for our best model which is in the PowerPoint presentation(Best Model was determined by the best AIC score). Here are the corresponding graphs for the proportions based on our best model which includes race, gender, and armed threat level.

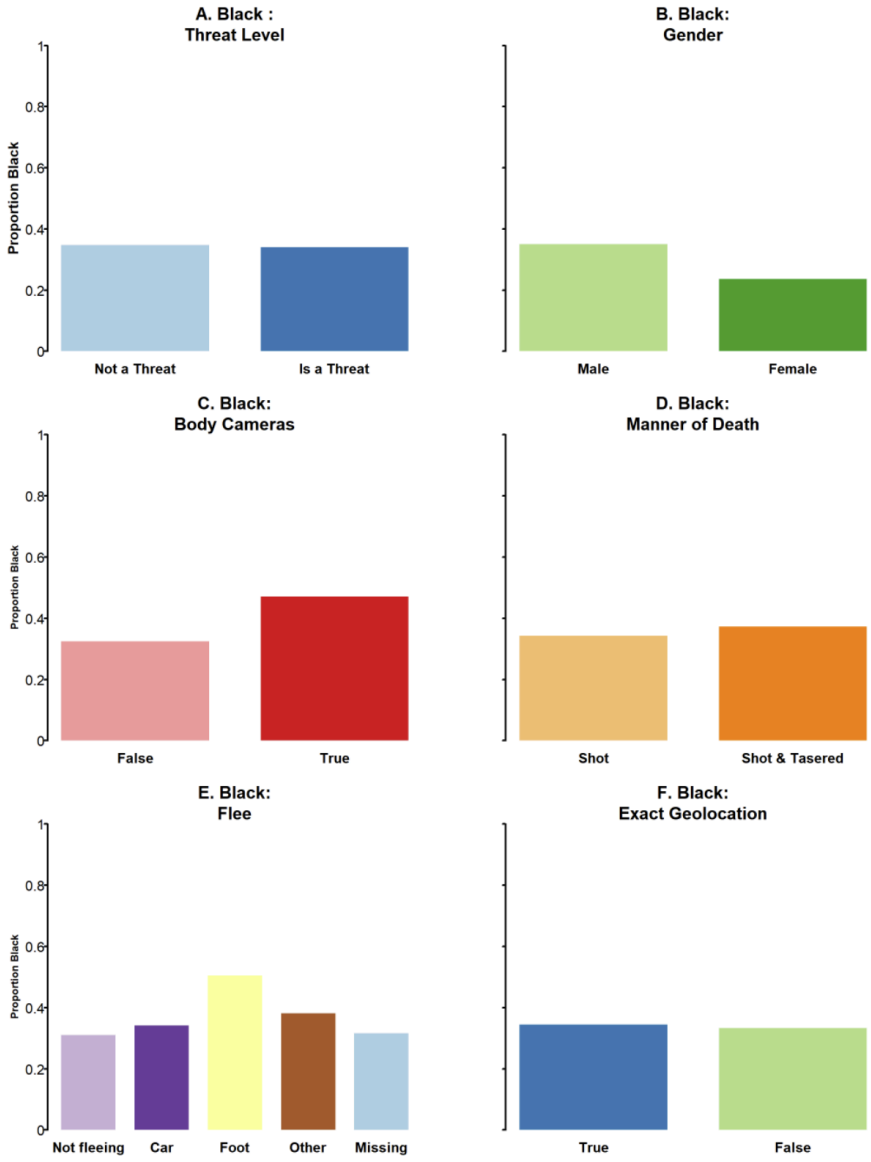


Based on the feedback received, we also wanted to show race as the outcome variable, (0 for white, 1 for black). Luckily, we already started making code before the presentation to see if race would be a better outcome variable, and here were some preliminary results:

Description: df [1 x 4]

Accuracy <dbl>	FalseNeg <dbl>	FalsePos <dbl>	PositivePred <dbl>
0.6823529	0.7081081	0.1210884	0.5482234

1 row



Essentially, we flipped the numerical variable we created, through data wrangling, with threat level in our best model. The model is still very poor and even has a worse positive predict rate. While I believe this is the right approach, we would definitely need to do more data wrangling to be able to guess the race of a fatal police shooting victim. Approaches to alleviating this issue would be to make the whole data set slightly more usable in the glm function because comparing 50 different states would be taxing on the CPU but also just finding the right variables because I think just the categorical data is not enough to create the best model.

Conclusion

Although we proved our alternate hypothesis (1% difference), it could also be seen as having no effect since it is only 1%. However, this is up to interpretation if we think about the Census Data.

Limitations:

First off there was a lack of numerical data (most of the variables had categorical values). For example, in the gender column, the value for a male would be “male” instead of a 0 or 1. Second, there was too much data wrangling to do. Lastly, there were misleading proportions (actual vs census data). This issue could’ve been avoided with a stratified random sample but that would be data wrangling on top of what is already given.

Future Directions:

Using a different outcome variable to create a better model. We started doing that with race as our outcome variable. There are also other categories instead of just race and gender that have an impact in causing black males to have a higher probability of being described with an attack threat level than white people. There were many other useful variables in the data set and we would like to fine-tune which combination is the most accurate.