




The Data Deities



Adam Ford, Allan Juarez, Carter Wunsch,
Joseph Strobel, Patrick Wenzel



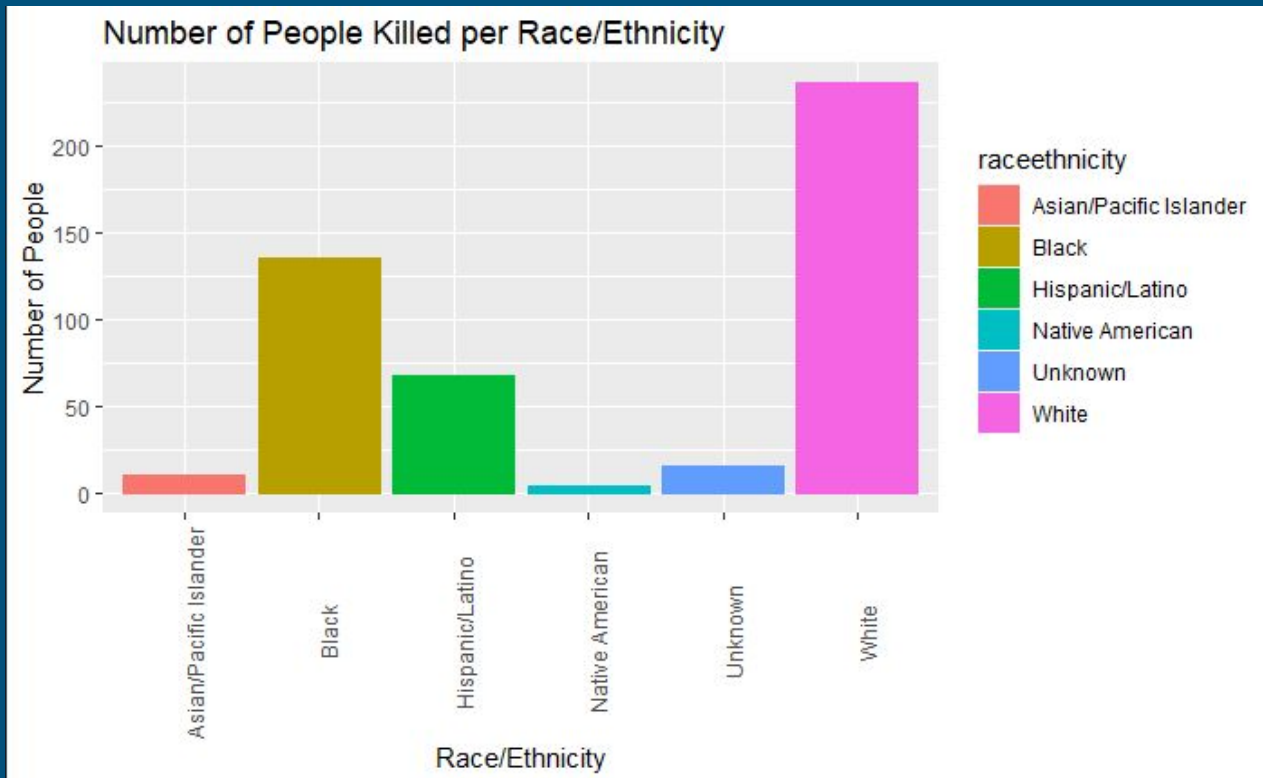
Background

- Police killings in America from 2015
- Why is this dataset important?

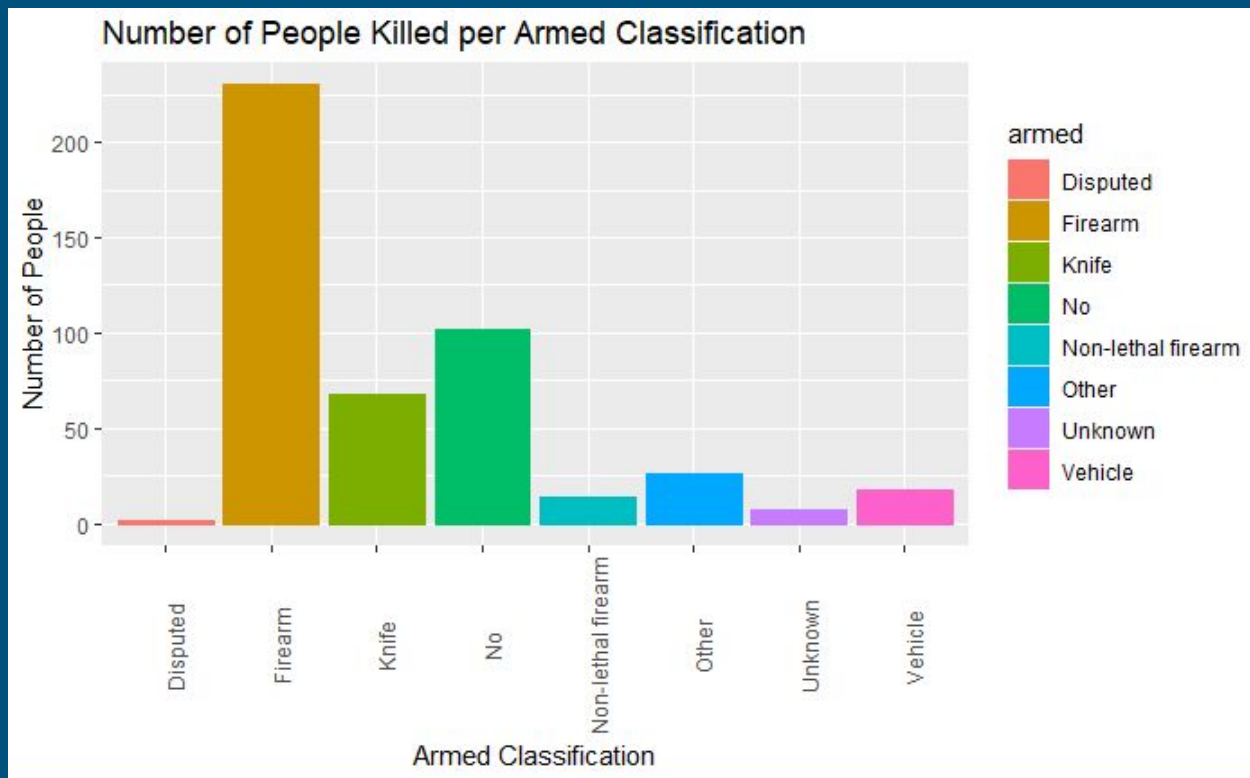
Background - Predictors

- armed: How/Whether the deceased was armed
- age: How old the deceased was
- raceethnicity: Race/ethnicity of the deceased
- pov: Tract-level poverty rate
- h_income: Tract-level median household income
- urate: Tract-level unemployment rate
- cause: Cause of death
- gender: Gender of the deceased
- college: Share of 25+ pop with BA or higher

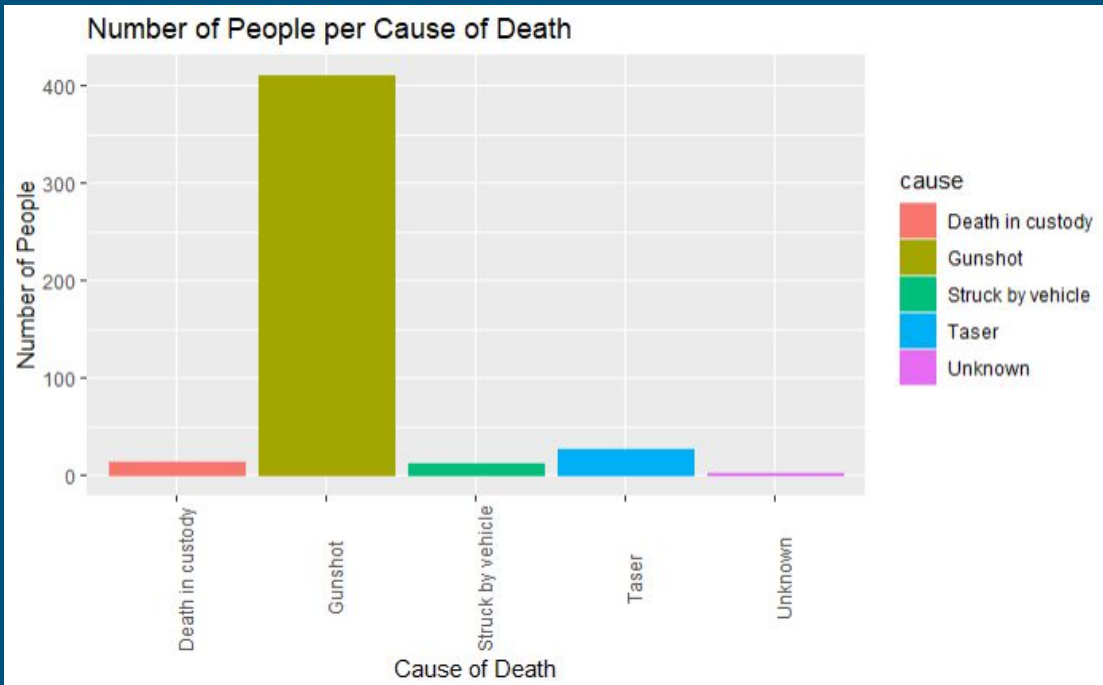
Exploratory Data Analysis (EDA)



EDA Cont.

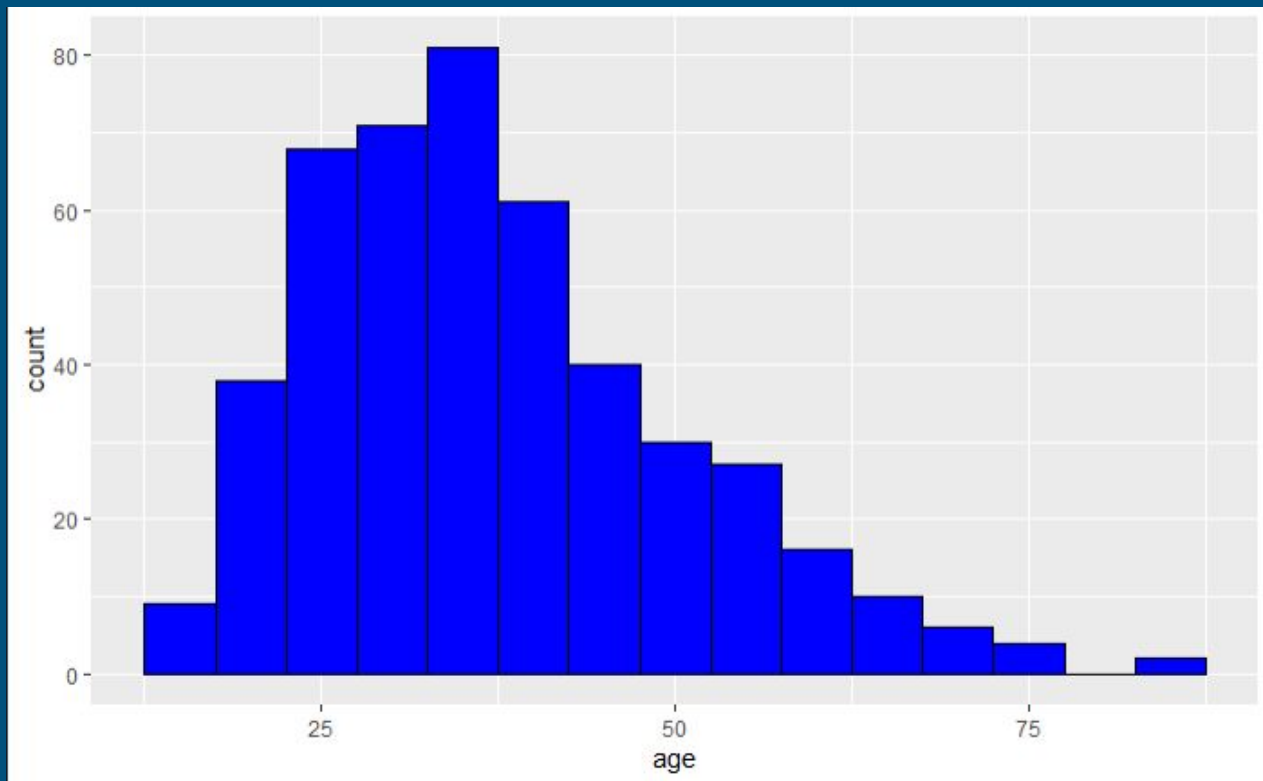


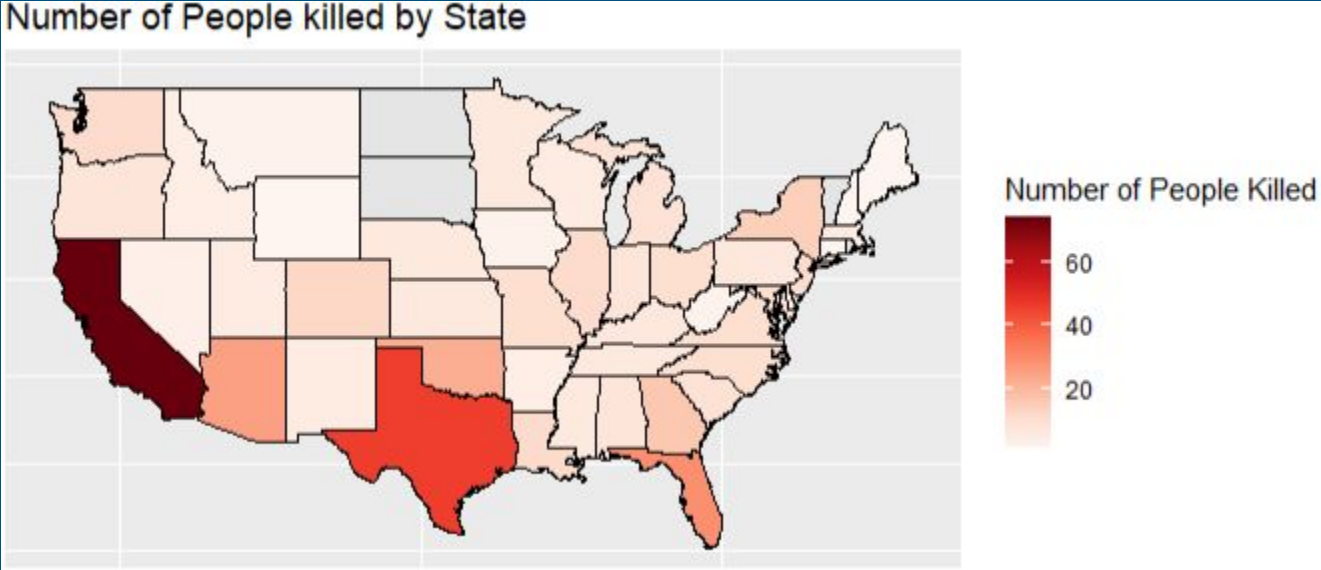
EDA Cont.



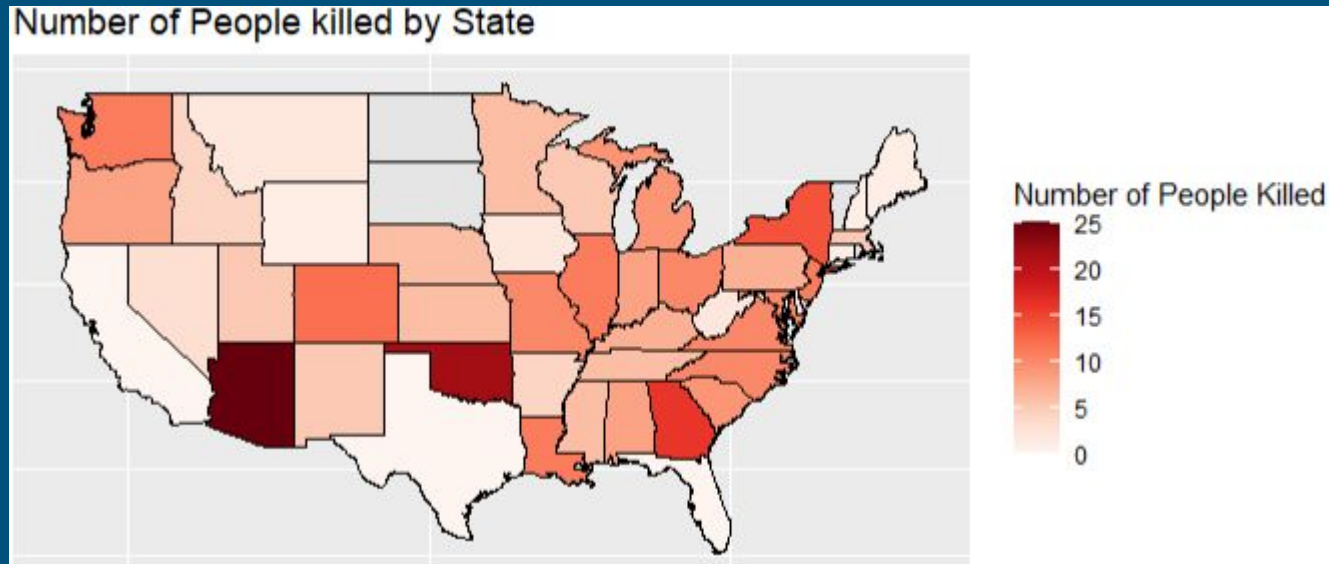
cause	numPeople
<chr>	<int>
Death in custody	14
Gunshot	411
Struck by vehicle	12
Taser	27
Unknown	3

EDA Cont.





EDA Cont.



Questions

1. Can you predict if the person killed by the police is white based on several of the given predictors?
2. Can you predict if the person was armed based on several of the given predictors?

Question 1 - Logistic Regression - Initial Runs

Using Personal Information:

(Age, Gender, Armed)

Classification Rate: 59.4%

	Non-White	White
Predicted Non-White	60	36
Predicted White	58	78

Using Locational Data as Interactive:

(Tract College Graduation Rate *
Unemployment Rate * Poverty Rate)

Classification Rate: 61.6%

	Non-White	White
Predicted Non-White	60	31
Predicted White	58	83

Note: There is not one error more harmful than another, ROC not used

Question 1 - Logistic Regression - Subset

The subset that minimizes AIC, BIC:

(Armed, Age, Tract College Graduation Rate * Unemployment Rate * Poverty Rate)

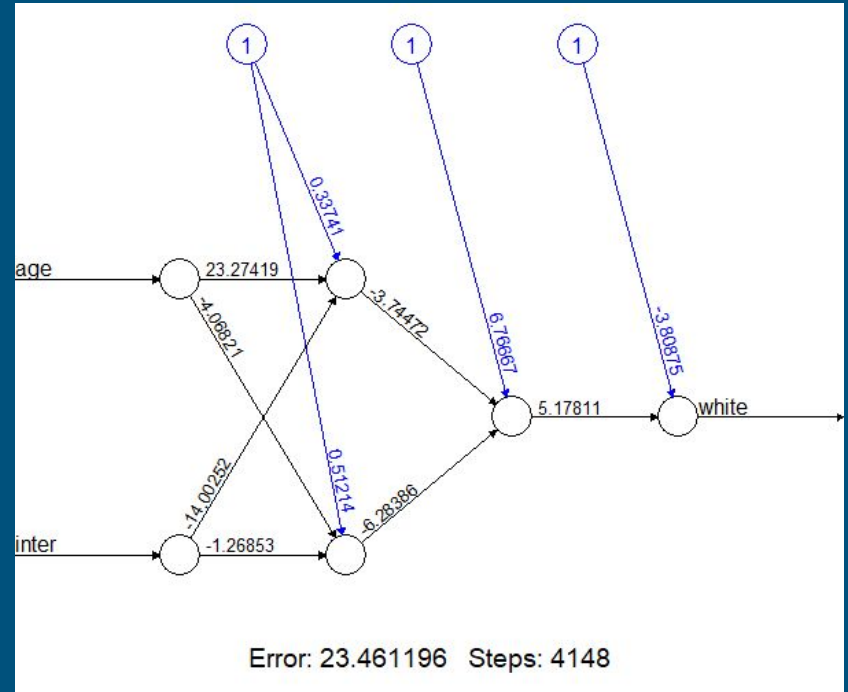
Classification Rate: 65.1%

	Non-White	White
Predicted Non-White	71	34
Predicted White	47	80

Question 1 - Neural Network

Used Armed, Age, Tract College
Graduation Rate * Unemployment Rate *
Poverty Rate for predictors

Classification Rate: 63.2%



Question 1 - QDA Exploration

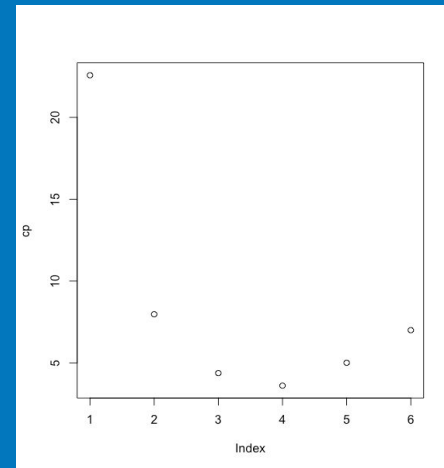
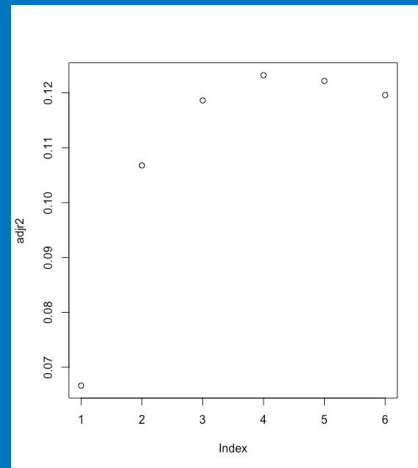
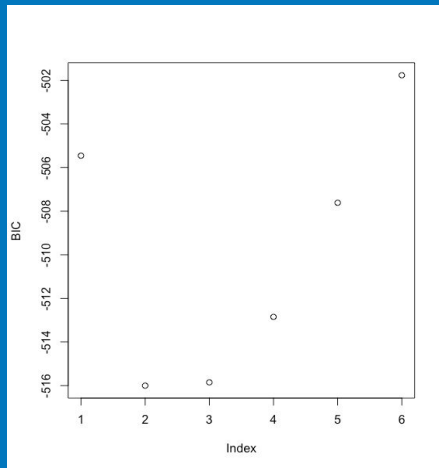
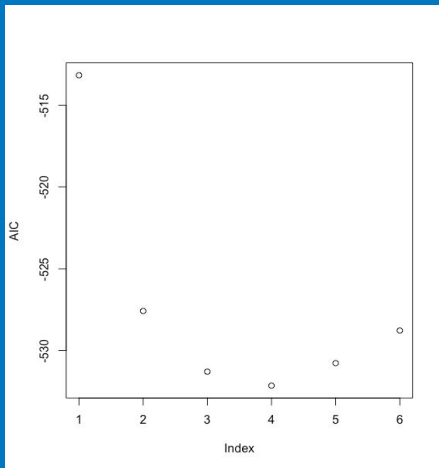
Began by choosing pairs of predictors I thought may be useful. (age, cause, pov, college, armed)

Classification Rate

	age	cause	pov	college	armed
gender	58.97	53.85	60.68	44.44	47
age		53.85	65.81	58.11	58.97
cause			62.39	52.99	52.99
pov				57.26	64.96
college					47.86

Question 1 - Forward subset selection

Selected from: age, cause, armed, pov, college, h_income



Question 1 - Forward subset selection

```
> coef(regfit, 3)
```

(Intercept)	age	pov	h_income
6.780493e-01	8.906533e-03	-1.303666e-02	-4.445995e-06

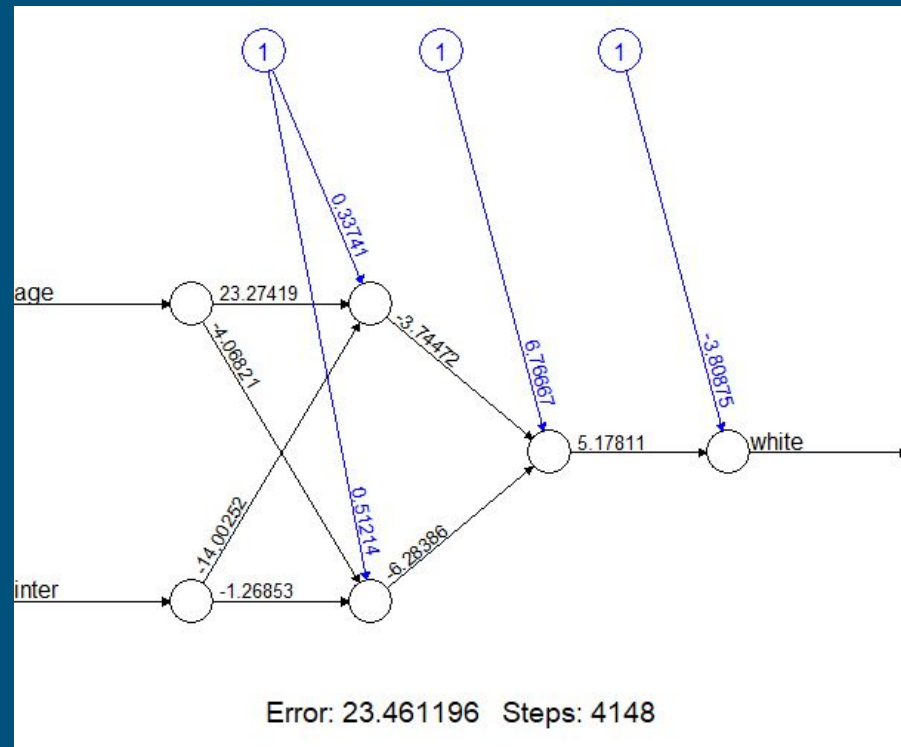
Classification rate: 58.1% on the test set

Confusion matrix:

	0	1
0	23	10
1	39	45

Question 1 - Neural Network

This ANN still only achieved a 63.2% classification rate.



Question 2 - GLM

- For our first approach we used the glm function to create a logistics regression model using predictors we thought were relevant to predicting whether the person was armed or not
- We choose age, race/ethnicity, poverty rate, income, unemployment, cause of death, gender, and college experience
- Overall misclassification rate: 0.225

Confusion matrix:

glm.pred	Unarmed	Armed
Unarmed	2	7
Armed	42	167

Question 2 LDA

- We also used the same predictors in the LDA model in order to compare misclassification rates of the two models
- Overall misclassification rate: 0.214
- Confusion matrix:

glm.pred	Unarmed	Armed
Unarmed	2	5
Armed	41	167

Forward Selection

- Fitted model with armed as the response
- We chose the same original predictors that we used in glm and lda
 - Based on the data provided, these predictors seemed most relevant
- We ended up choosing our final model based on the smallest BIC and smallest cp which had a model size of 3
- The best predictors were age, race/ethnicity, and cause of death

AIC_min	BIC_min	adjr2_max	cp_min
3	1	7	3

Question 2 GLM using Stepwise Selection

- Next we went back and trained another logistic regression model using the three predictors found from stepwise selection
- Overall misclassification rate: 0.193

Confusion matrix:

glm.pred	Unarmed	Armed
Unarmed	0	2
Armed	40	176

Question 2 LDA using Subset Selection

- We also went back and trained an LDA model using the three predictors we got from the forward stepwise selection
- Overall misclassification rate: 0.185

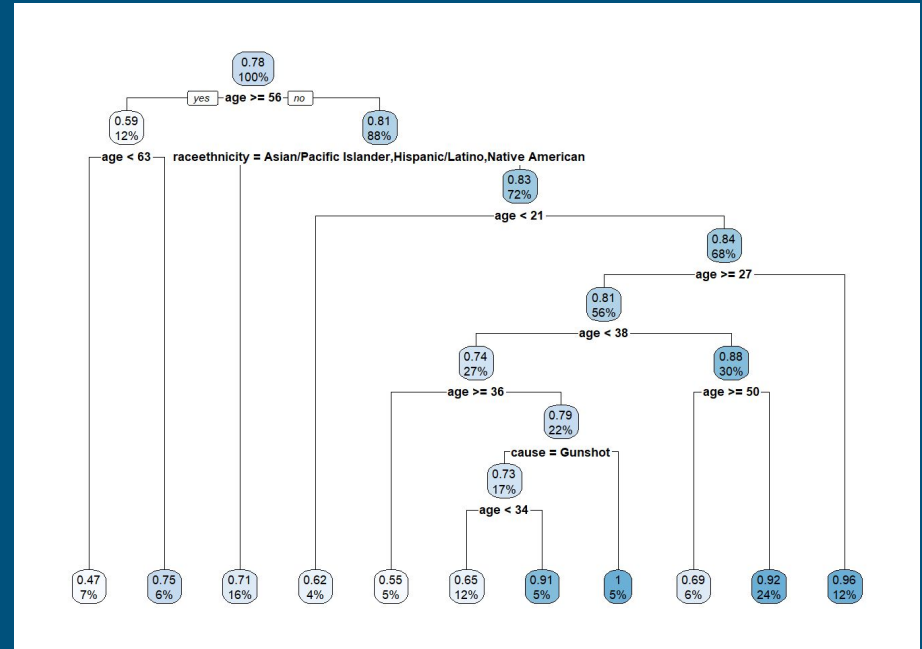
glm.pred	Unarmed	Armed
Unarmed	0	4
Armed	36	176

Recursive Partitioning Explanation

- Recursive Partitioning creates a decision tree. It is known to be recursive since it splits each of its subset members indefinite times until the terminal nodes are reached
- Pros:
 - R function provides intuitive models
 - Various methods of preventing overfitting
- Cons:
 - Overfitting tends to be an issue
 - Doesn't work well for continuous variables

RPart findings and conclusions

- At each node, there is a condition that determines which way you travel down the tree
- When you reach the end of the tree, the terminal nodes tell you what the chances are of that person being armed
- Considering there are the most nodes with age as the condition, we believe that age is a significant predictor of whether the victim was armed



Conclusion

- Can you predict if the person killed by the police is white based on several of the given predictors?
 - No, there was no conclusive method of classifying if a person is white based on our predictors.
 - Only as high as 65% classification rate
- In general, we wish we had a dataset that included non-fatal police encounters to better understand what makes an encounter deadly.

Conclusion - cont.

Can you predict if a person was armed at the time of their death based on several of the given predictors?

- Age, race/ethnicity, and cause of death were the most significant predictors of a person being armed and we can decently guess if the person was armed based on these predictors

What is the best model for predicting if a person was armed at the time of their death?

- LDA was the best model for prediction because it had the lowest misclassification rates

References

- N/A, dmill, andrewflowers N/A, and Andrei Scheinkman. *Police_killings.csv*. GitHub: Fivethirtyeight, 3 June 2015. CSV.
- Viner, Katharine, Lee Glendinning, and Matt Sullivan. "About the Counted: Why and How the Guardian Is Counting US Police Killings." 2 June 2015. Web. 30 Apr. 2021.
- Weessies, Kathleen. "Finding Census Tract Data: About Census Tracts." 23 Feb. 2010. Web. 30 Apr. 2021.



Thank you!

Any Questions?