US Fatal Police Shootings Data Analysis

Xufan Wang, Yifei Song, Zhirui Li

Data Resources

The main dataset we choose is the Police Shootings in the US from https://www.washingtonpost.com/graphics/investigations/police-shootings-database/.

This database contains records of every fatal shooting in the United States by a police officer in the line of duty since Jan. 1, 2015. It is updated regularly as fatal shootings are reported and as facts emerge about individual cases.

Note that this does not include civilians killed in police custody, fatal shootings by off-duty officers, and non-shooting deaths.

The data was gathered via law enforcement websites, local news reports, and social media.

The data set includes data such as: name, date, manner of death, whether the person was armed, the age, gender, and race of the person, the city and state where the shooting took place, whether the person showed signs of mental illness, the threat level of the incident, whether the person was fleeing, and whether the officer in question had a body camera.

Research Question

1. What variables have the most effect to the case number of fatal police shooting
2. Build a model to predict whether fatal police shooting victims' race is Black

Exploratory Data Analysis

Overview

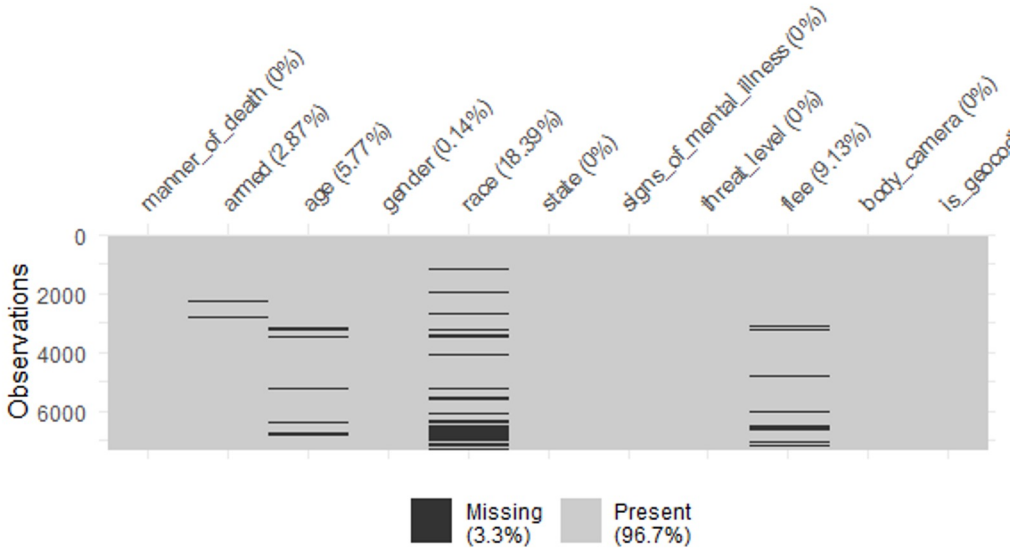
Dataset statistics

Number of variables	17
Number of observations	7291
Missing cells	4428
Missing cells (%)	3.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	818.9 KiB
Average record size in memory	115.0 B

Variable types

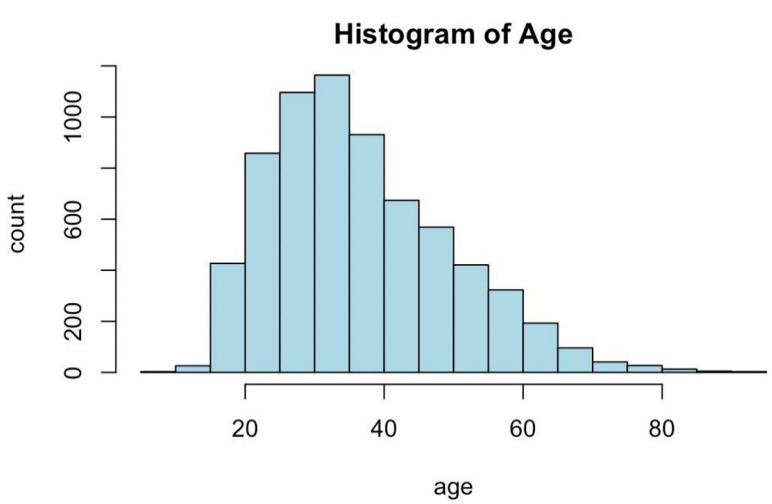
Numeric	4
Categorical	10
Boolean	3

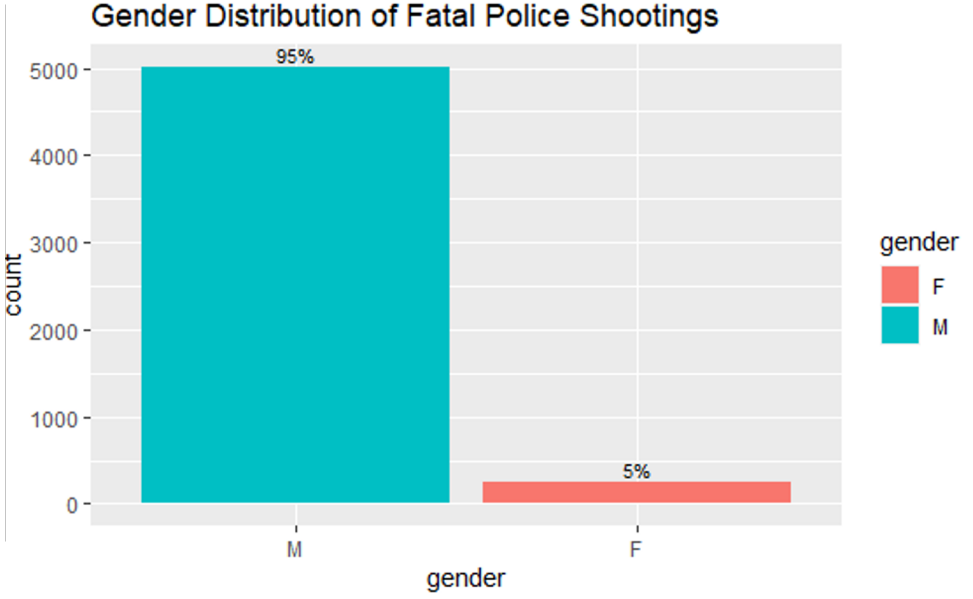
Missing Value



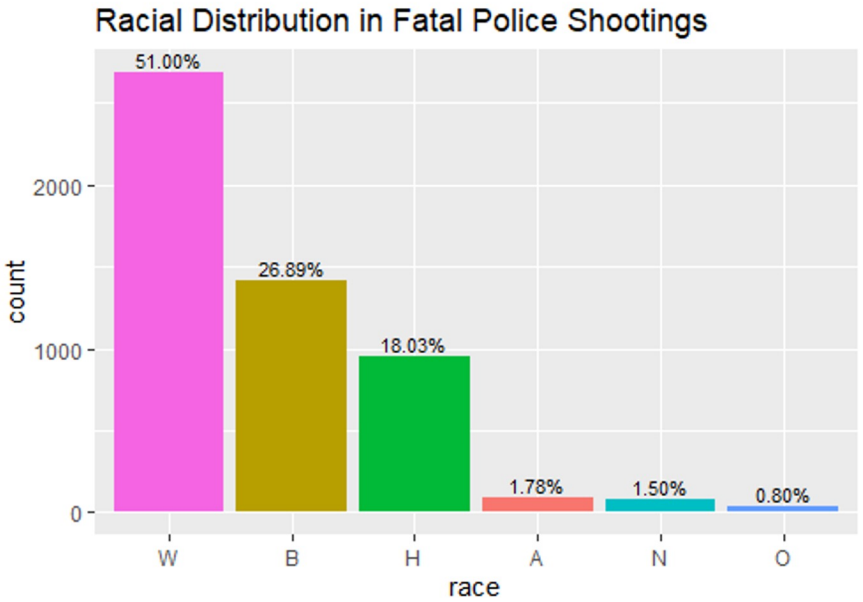
Data Visualizations

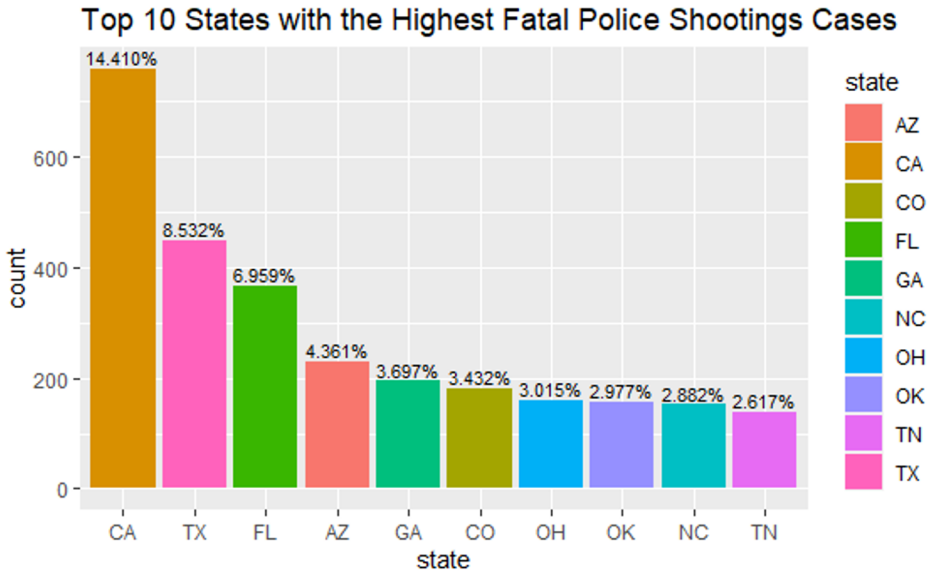
Age and Gender Distribution



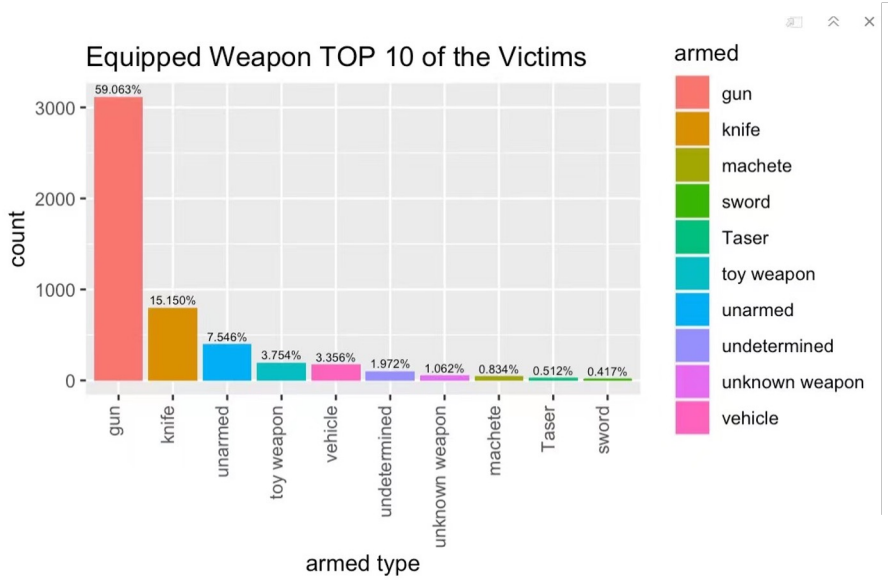


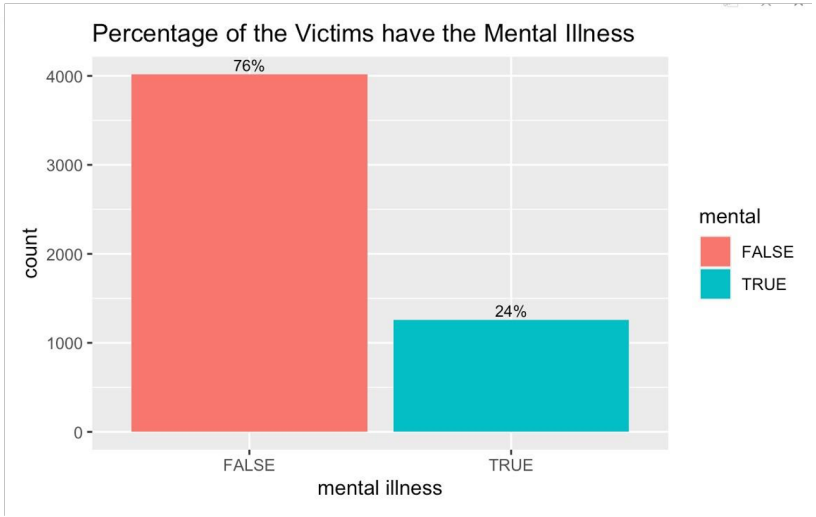
Racial Distribution and States with Top 10 Cases



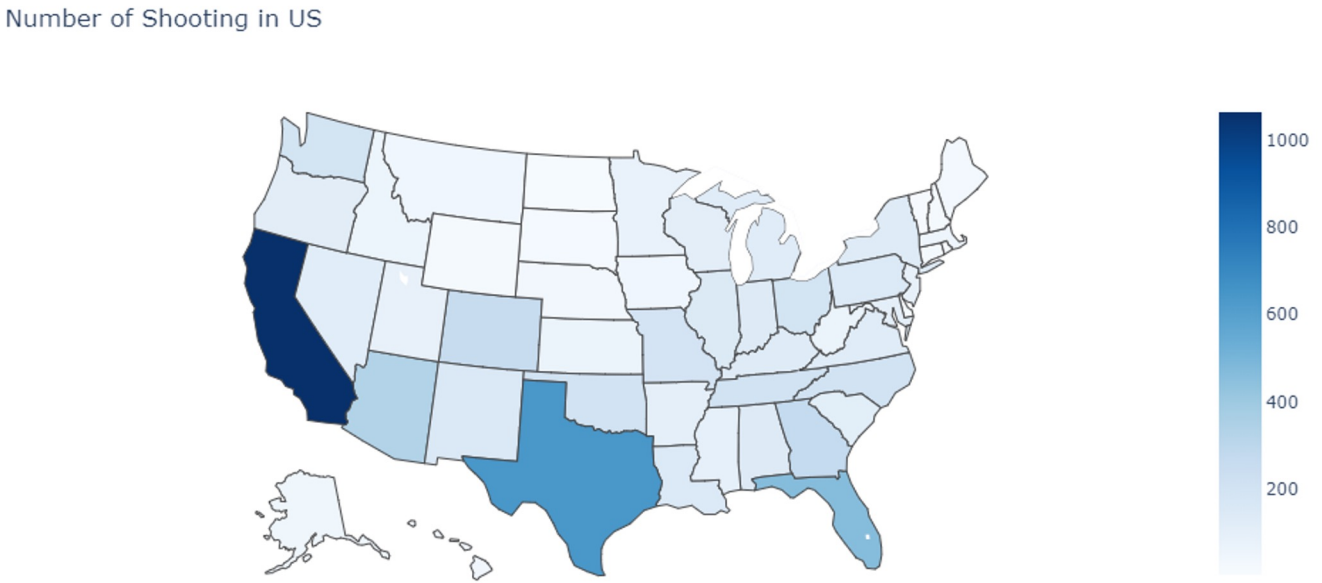


Top 10 Armed Type and Mental Illness





Geographic graph of shooting cases in US



Further Research

with Additional Data

Data Source

There are altogether three additional datasets. These are US census data on poverty rate, high school graduation rate, median household income. These dataset can be found from the following link:

https://www.kaggle.com/datasets/kwullum/fatal-police-shootings-in-the-us?select=MedianHouseholdIncome2015.csv

Overview

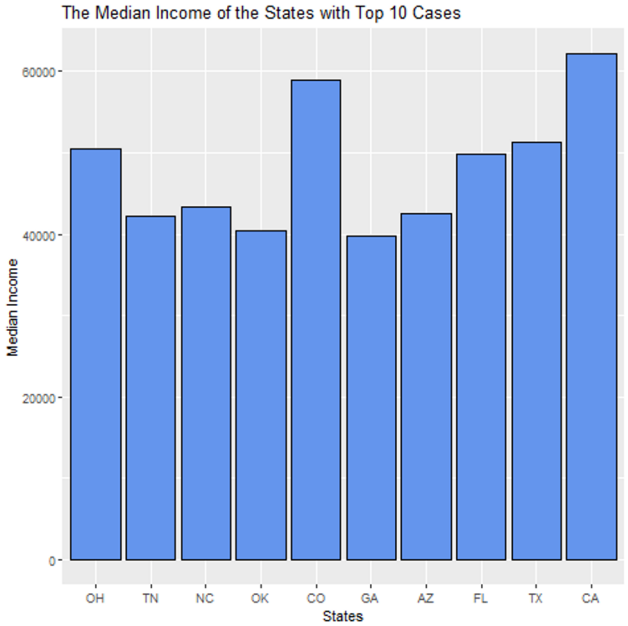
Dataset statistics

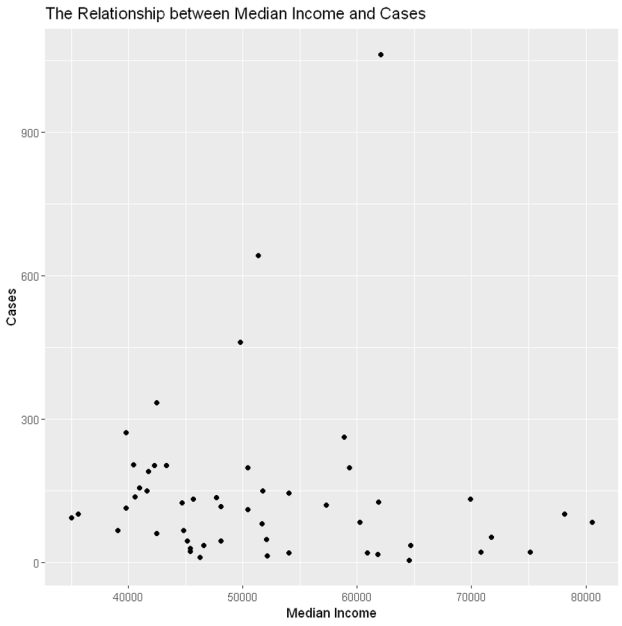
Number of variables	3
Number of observations	29329
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	687.5 KiB
Average record size in memory	24.0 B

Variable types

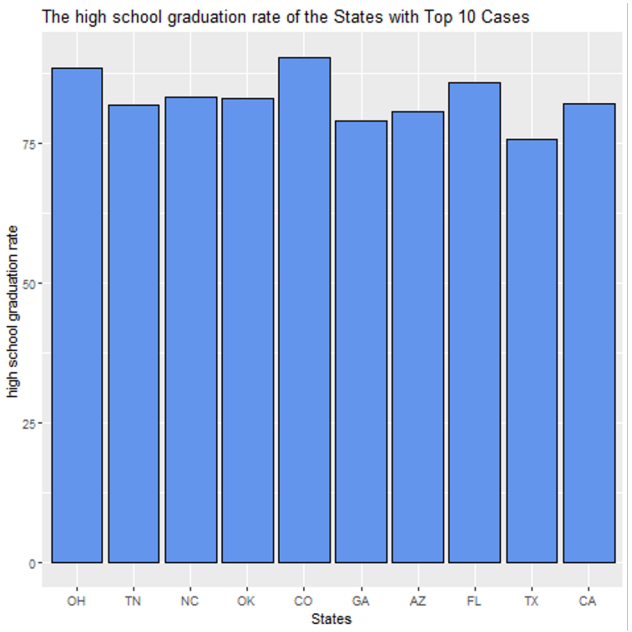
Categorical	3

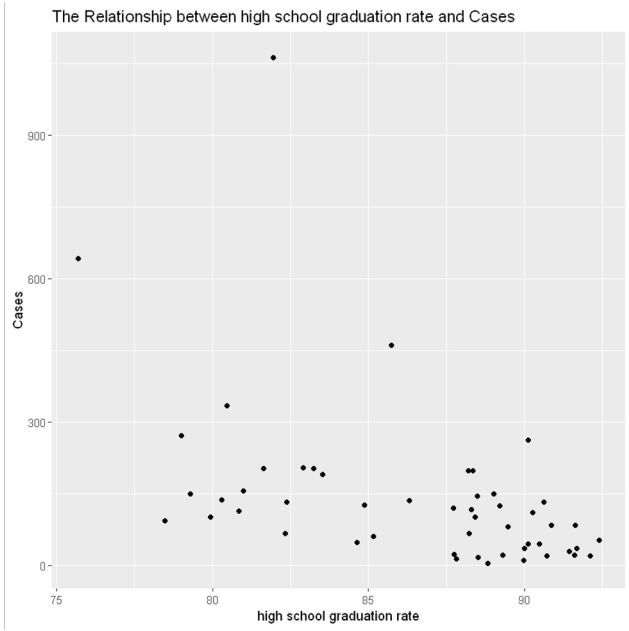
Median Income



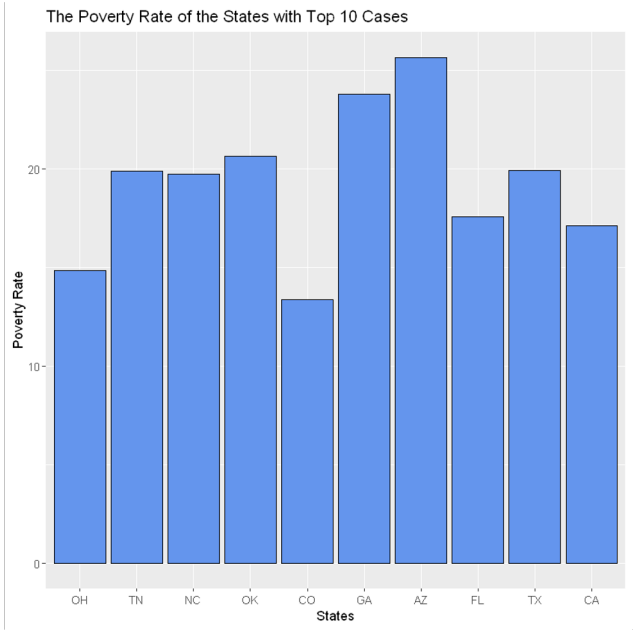


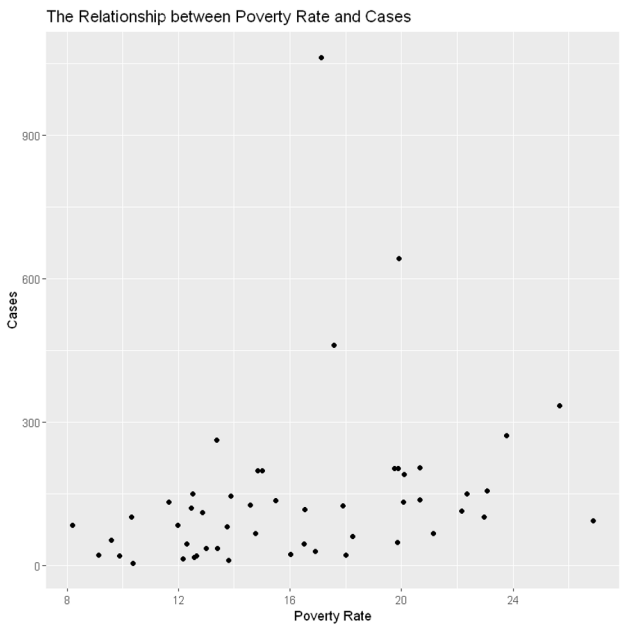
High School Graduation Rate





Poverty Rate





Logistic Regression

Logistic Regression Model

1. If race is black, encoded as 1. Otherwise, encoded as 0.
2. Convert all categorical variables to factors.
3. Standardize all continuous variables to prevent multicollinearity and help the model to converge faster.
4. Split the whole dataset into training(80%) and testing(20%).
5. Mixed effects logistic regression model and logistic regression model chosen by forward stepwise regression without random effects.

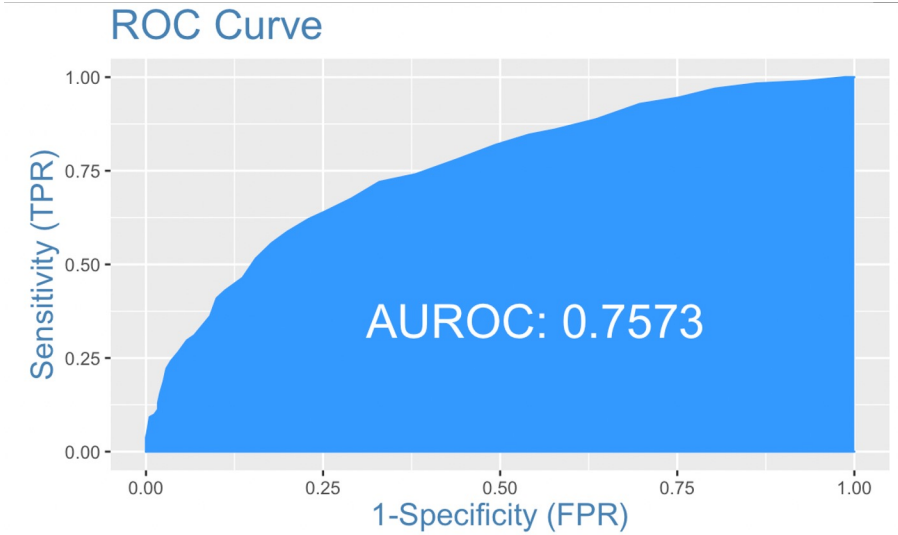
Mixed Effects Logistic Regression

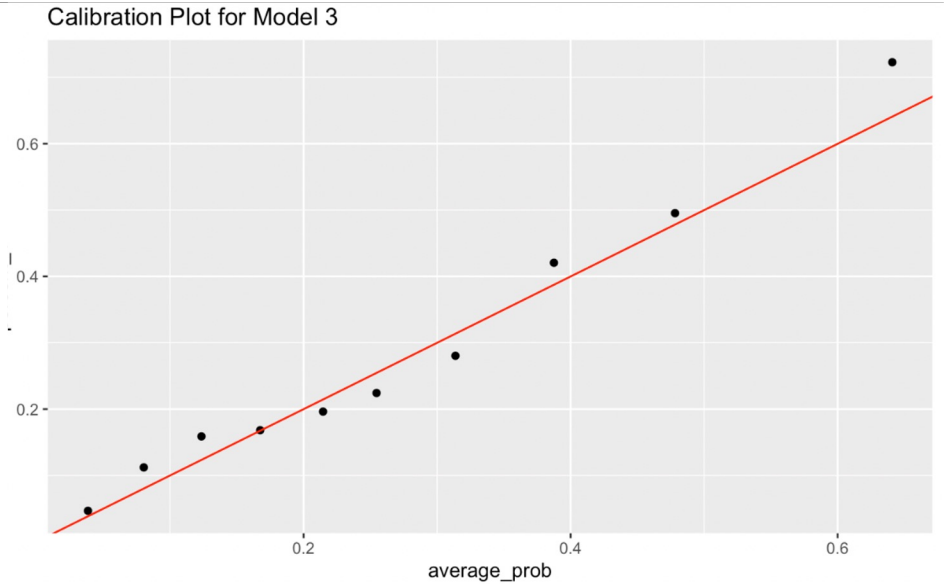
```
model1 = glmer(race ~ manner_of_death + armed + gender + signs_of_mental_illness
+ threat_level + flee + body_camera + is_geocoding_exact + age + (1|state), data
= fatal, family = 'binomial', control=glmerControl(optimizer='optimx',
optCtrl=list(method='nlminb'), nAGQ=9))
```

```
model2 = glmer(race ~ gender + signs_of_mental_illness  + flee + body_camera +
age + (1|state), data = fatal, family = 'binomial',
control=glmerControl(optimizer='optimx', optCtrl=list(method='nlminb'), nAGQ=9))
```

```
model3 = glmer(race ~ gender + signs_of_mental_illness + flee + body_camera + age
+ (1|state) + gender:flee, data = fatal, family = 'binomial',
control=glmerControl(optimizer='optimx', optCtrl=list(method='nlminb'), nAGQ=9))
```

Mixed Effects Logistic Regression





* Graphs are based on test set.

Mixed Effects Logistic Regression

Top 10 States:

	grpvar <chr>	term <fctr>	grp <fctr>	condval <dbl>	condsd <dbl>
8	state	(Intercept)	DC	2.597200	0.5367490
21	state	(Intercept)	MD	1.856567	0.2368605
19	state	(Intercept)	LA	1.699653	0.2036552
35	state	(Intercept)	NY	1.403079	0.2141763
32	state	(Intercept)	NJ	1.363625	0.2771099
15	state	(Intercept)	IL	1.260613	0.1989739
46	state	(Intercept)	VA	1.150454	0.2106795
11	state	(Intercept)	GA	1.096554	0.1516108
25	state	(Intercept)	MO	1.040313	0.1836298
39	state	(Intercept)	PA	1.035038	0.2036082

Bottom 10 States:

	grpvar <chr>	term <fctr>	grp <fctr>	condval <dbl>	condsd <dbl>
33	state	(Intercept)	NM	-2.116190	0.4955930
27	state	(Intercept)	MT	-1.836926	0.7124800
14	state	(Intercept)	ID	-1.626104	0.6022155
29	state	(Intercept)	ND	-1.308097	0.8025432
42	state	(Intercept)	SD	-1.265575	0.8023424
4	state	(Intercept)	AZ	-1.248993	0.2397495
38	state	(Intercept)	OR	-1.192308	0.4081365
51	state	(Intercept)	WY	-1.188253	0.8121211
12	state	(Intercept)	HI	-1.175502	0.6467943
31	state	(Intercept)	NH	-1.144006	0.8166288

Mixed Effects Logistic Regression

```r
DC_odds = exp(2.5972)
MD_odds = exp(1.856567)
NM_odds = exp(-2.116190)
DC_odds
MD_odds
NM_odds
```

```
[1] 13.42609
[1] 6.401722
[1] 0.1204898
```

```r
comparison = 13.42609/0.1204898
comparison
```

```
[1] 111.4293
```

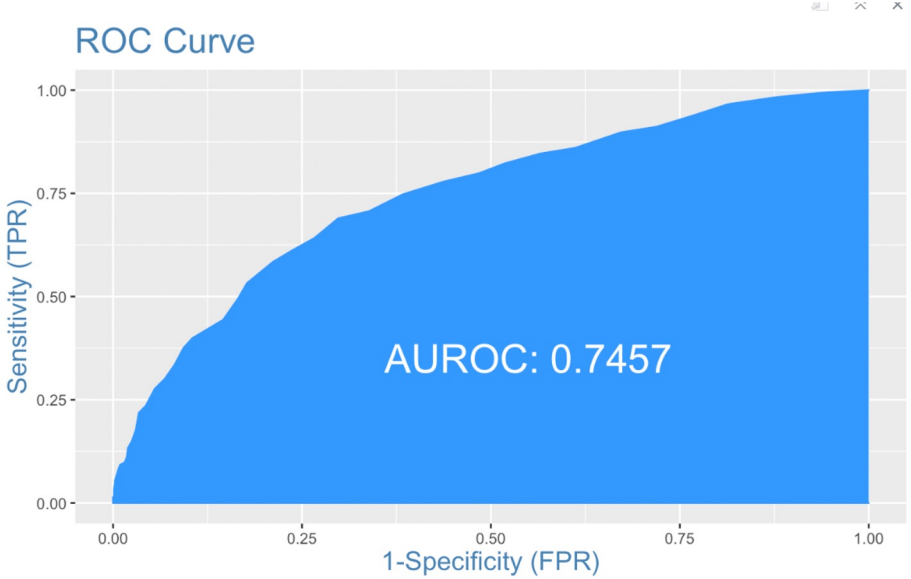
```r
probability_DC = DC_odds/(1+DC_odds)
probability_MD = MD_odds/(1+MD_odds)
probability_NM = NM_odds/(1+NM_odds)
probability_DC
probability_MD
probability_NM
```

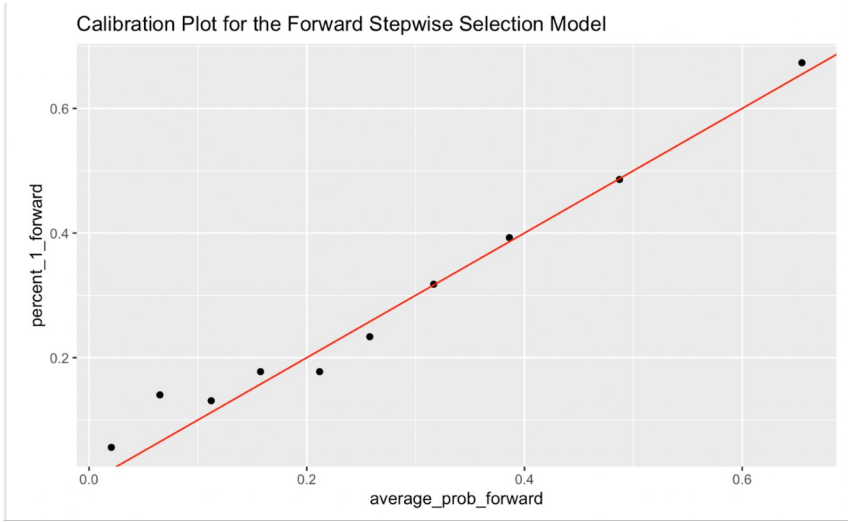
```
[1] 0.9306812
[1] 0.8648963
[1] 0.1075332
```

Forward Stepwise Selection without Random Effects

```
Step:  AIC=4178.37
race ~ state + age + signs_of_mental_illness + body_camera +
    flee + gender + threat_level
```

Forward Stepwise Selection without Random Effects





* Graphs are based on test set.

Forward Stepwise Selection without Random Effects

Top 10 States:

coefficient <dbl>	state <chr>
5.240528	stateDC
3.190929	stateMD
3.035911	stateLA
3.016233	stateDE
2.928156	stateRI
2.674353	stateNJ
2.562339	stateNY
2.524541	stateIL
2.391383	stateVA
2.385193	stateGA

Bottom 10 States:

coefficient <dbl>	state <chr>
-14.3366207	stateND
-14.0472566	stateMT
-13.8853277	stateVT
-13.7946228	stateWY
-13.6860520	stateHI
-13.6826684	stateSD
-13.5914421	stateNH
-3.3143445	(Intercept)
-1.9408092	stateNM
-0.9729923	stateID

Forward Stepwise Selection without Random Effects

```r
forward_DC = exp(-3.31434 + 5.240528)
forward_MD = exp(-3.31434 + 3.190929)
forward_ND = exp(-3.31434 - 14.3366207)
forward_DC
forward_MD
forward_ND
```

```
[1] 6.863297
[1] 0.8839003
[1] 2.159162e-08
```

```r
comparison_forward = 6.863297/2.159162e-08
comparison_forward
```

```
[1] 317868553
```

```r
probability_DC_forward = forward_DC/(1+forward_DC)
probability_MD_forward = forward_MD/(1+forward_MD)
probability_ND_forward = forward_ND/(1+forward_ND)
probability_DC_forward
probability_MD_forward
probability_ND_forward
```

```
[1] 0.8728269
[1] 0.4691863
[1] 2.159162e-08
```

Model Comparison on Test Set

	precision <chr>	recall <chr>	accuracy <chr>	AUROC <chr>
mixed effects model	66.4%	29.69%	76.30%	75.7%
model using FSS	62.9%	30.03%	75.64%	74.6%

Contributions

Xufan Wang, Yifei Song: EDA and Analysis

Zhirui Li: Logistic Regression Model

Thank you for watching!