Factors Related To Fatal Police Shootings in US

Team 2A
Cheng Chen | Jaya Nagesh | Mingwei Li | Selma Sentissi El Idrissi | Shamika Kalwe

Executive Summary

Project Proposal: Examining the various factors that are related to police shootings in the US

Goal Description:

- To analyze and visualise relationships between the independent variables (like gender, location, arms, age etc.) and the shooting incident. We can also explore incidents at different granularities like city or state or the arm used.
- Identify correlation between independent variables (if any).
- To try and predict (using ML if possible) some dependent variables in the dataset based on the independent variables.
- Try and combine with other datasets to draw deeper conclusions, for e.g. Population dataset (new addition)

Motivation

- Police brutality has been a raising concern across the world and this dataset about US Police Shootings seemed like a good place to start and take a deep dive into factors related to these shootings
- Further we wanted to verify if racial discrimination that the police has been accused of is reflected in actual data
- This dataset had scope to bring in new datasets like population metrics that would aid in deeper analysis
- Also, the scope to implement ML models was wide for this dataset given the multiple variables in it.

About the Data

> fatal-police-shootings-data.csv (486.61 KB)

Detail Compact Column

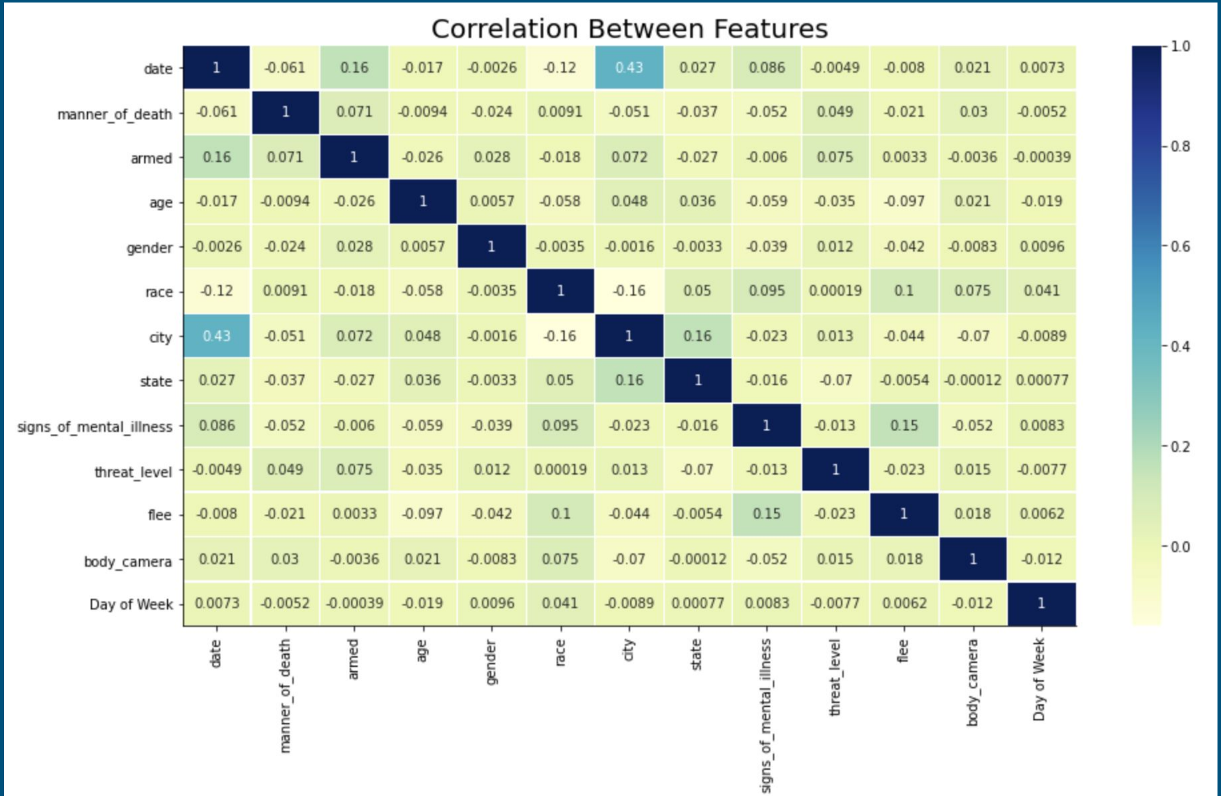
id	name	date	manner_of...	armed	age	gender	race	city	state	signs_of_...	threat_level	flee	body_cam...
3	Tim Elliot	2015-01-02	shot	gun	53	M	A	Shelton	WA	True	attack	Not fleeing	False
4	Lewis Lee Lembke	2015-01-02	shot	gun	47	M	W	Aloha	OR	False	attack	Not fleeing	False
5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23	M	H	Wichita	KS	False	other	Not fleeing	False

- Source: Kaggle
- Date range: 2-Jan-2015 - 16-Jun-2020
- 5416 rows, 14 columns
  - Contained null values
- Both categorical and numeric data
  - data types - string, date, boolean, integer
- Required moderate level of data cleaning prior to ML

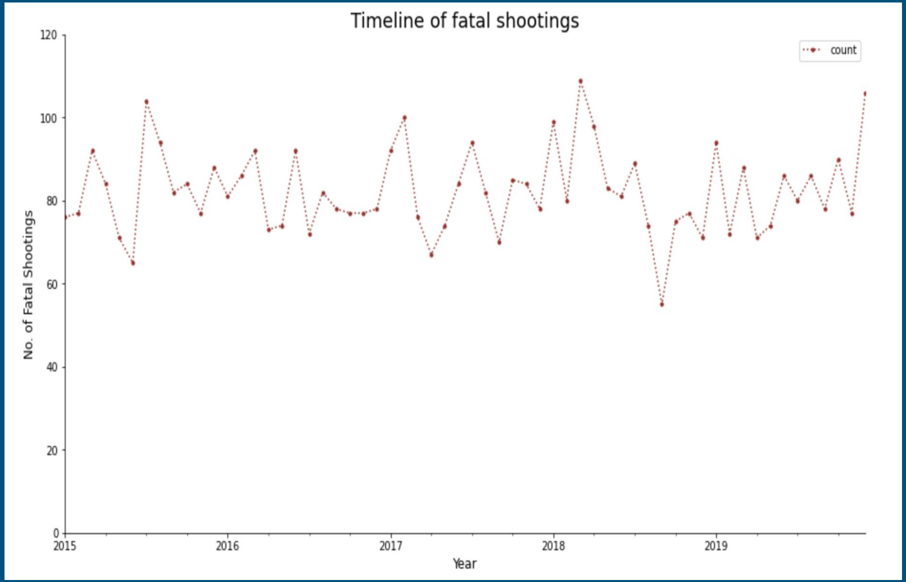
Exploratory Questions

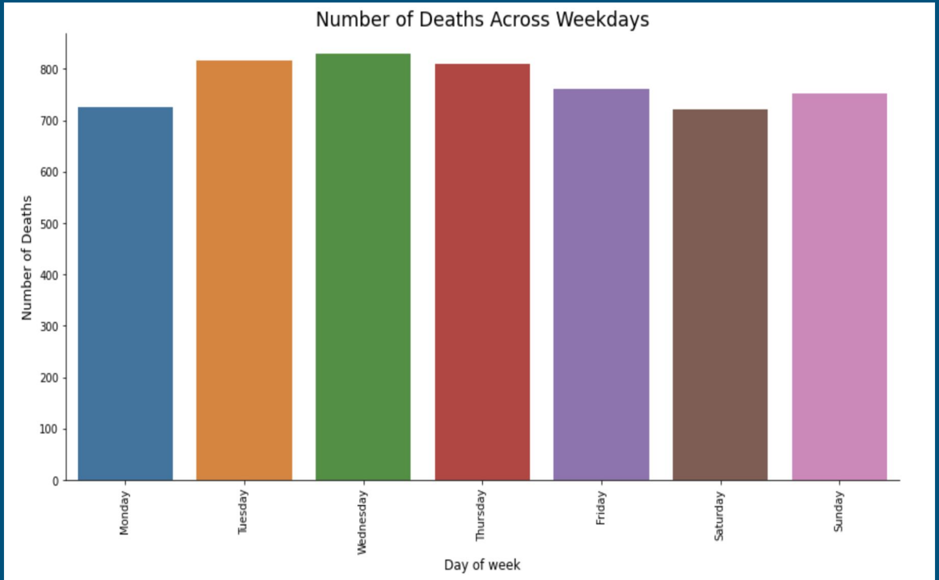
- What is the relationship between variables？
- Which state has the most deaths by a police officer?
- Which city has the most fatal shootings?
- What day has the highest deaths across the state level?
- What has been the timeline of these fatal shootings?
- How old are most of the victims?
- Did they flee at the time of shootings?
- What arms did the victims possess?
- What is the racial profile of the victims?

Correlation matrix plot



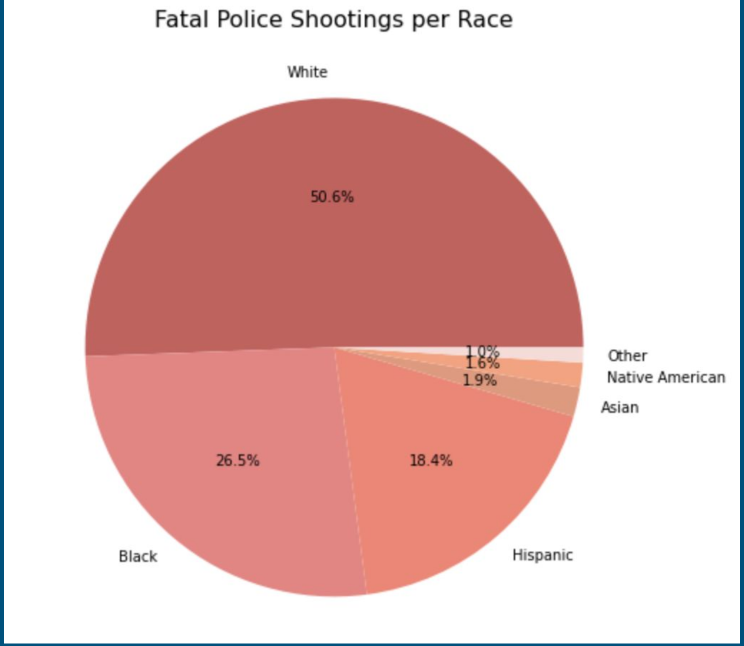
Exploratory Graphs

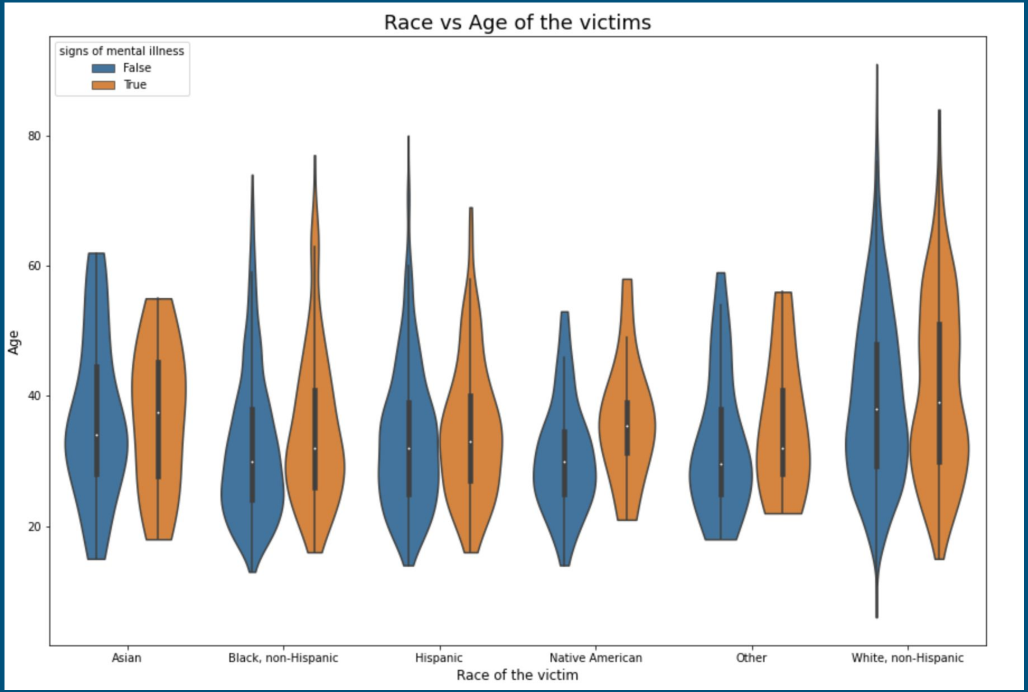




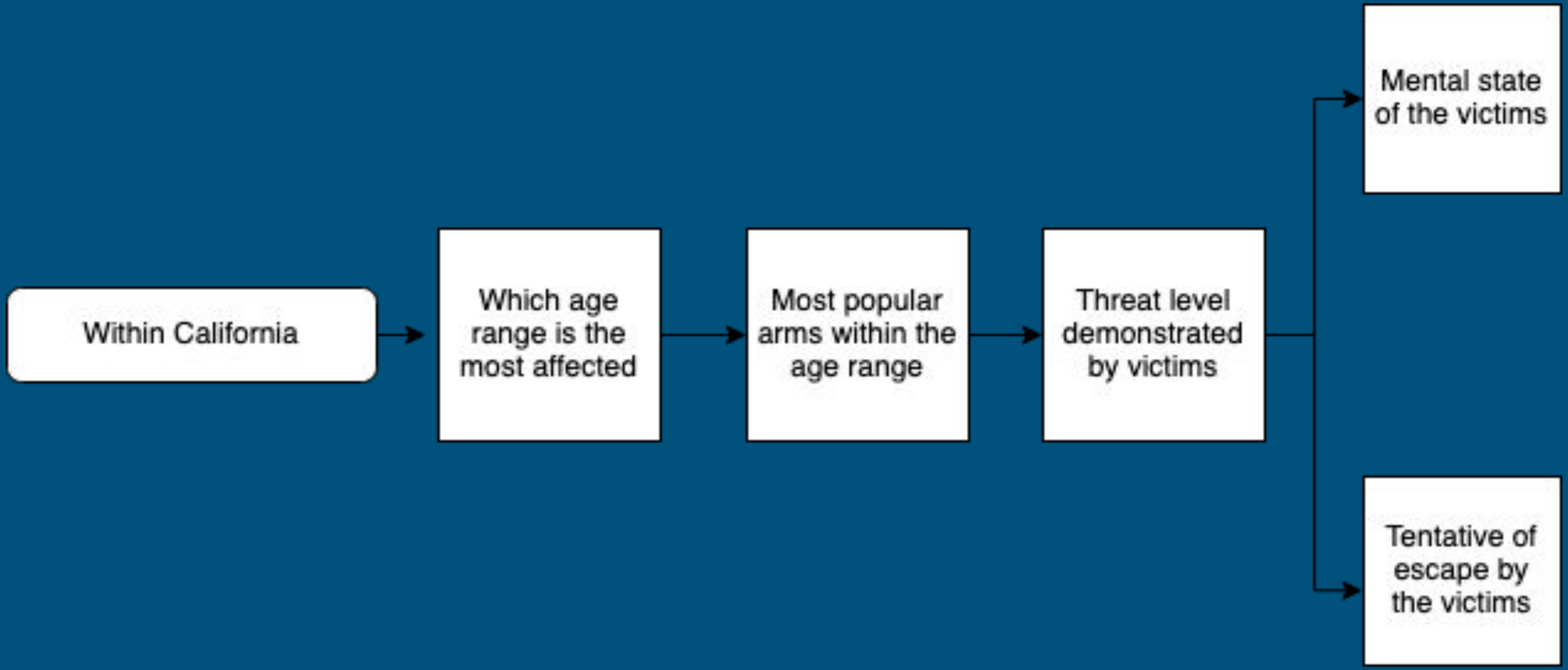
Note: The number for 2020 is only until June

Exploratory Graphs

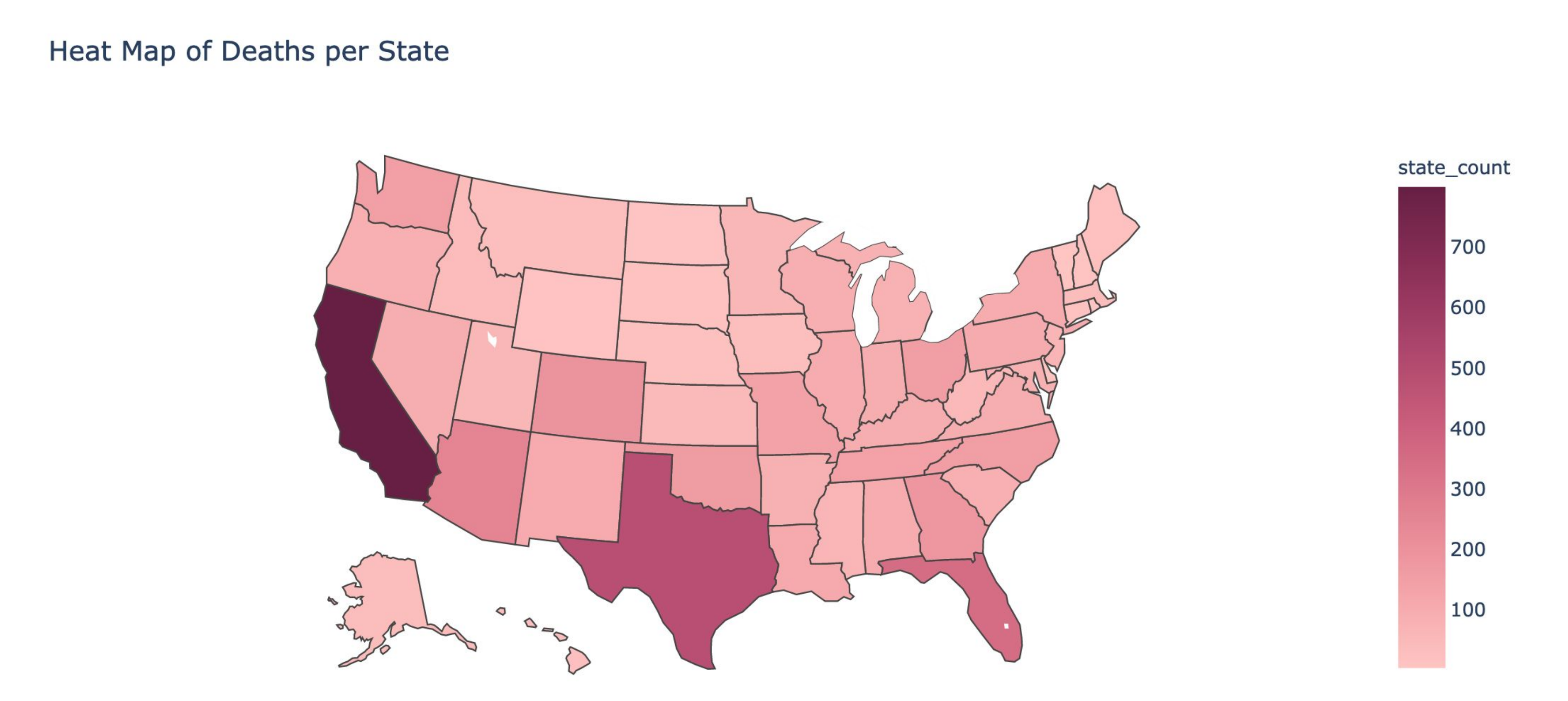




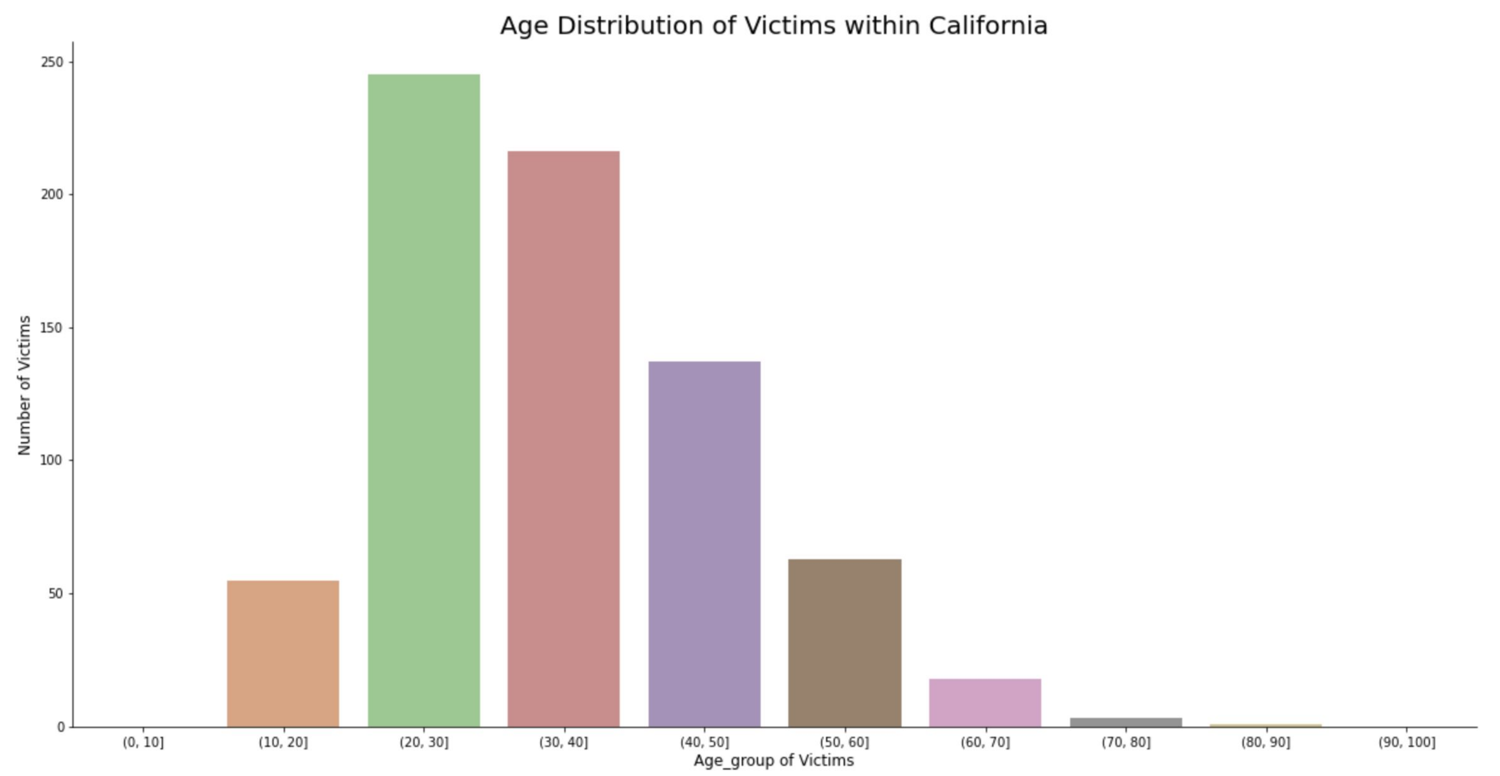
Introduction to Storyline



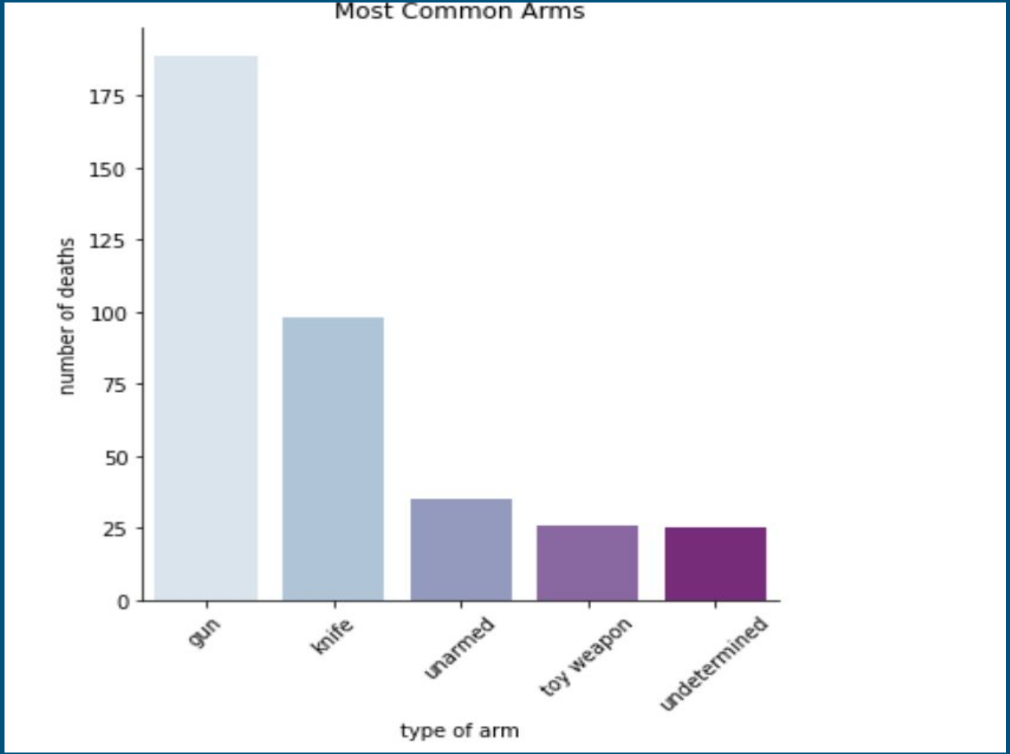
1. Which State has the most Deaths by Fatal Police Shootings?



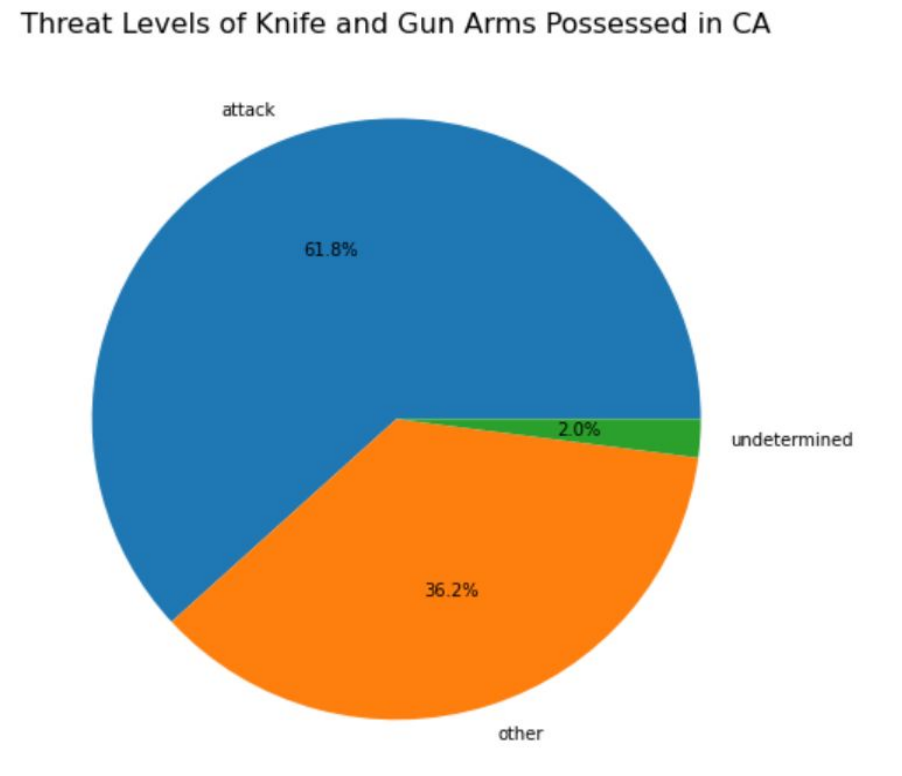
2. How old are most of the victims?



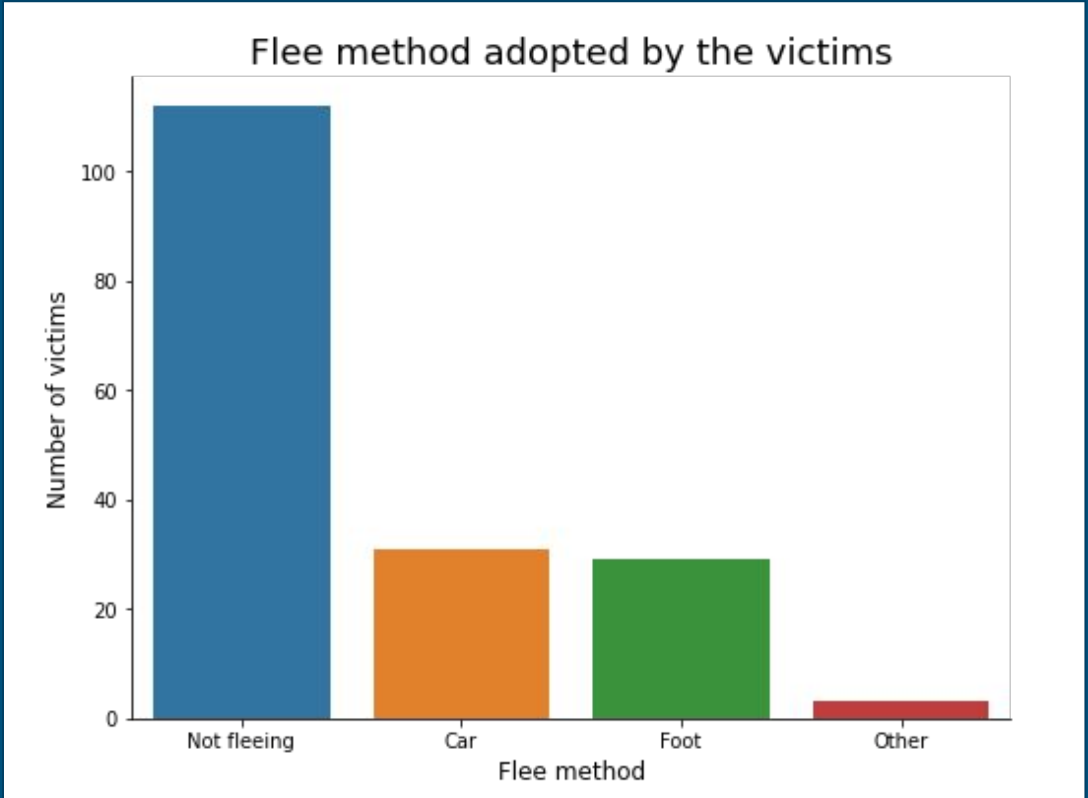
3. What were the most common arms possessed?



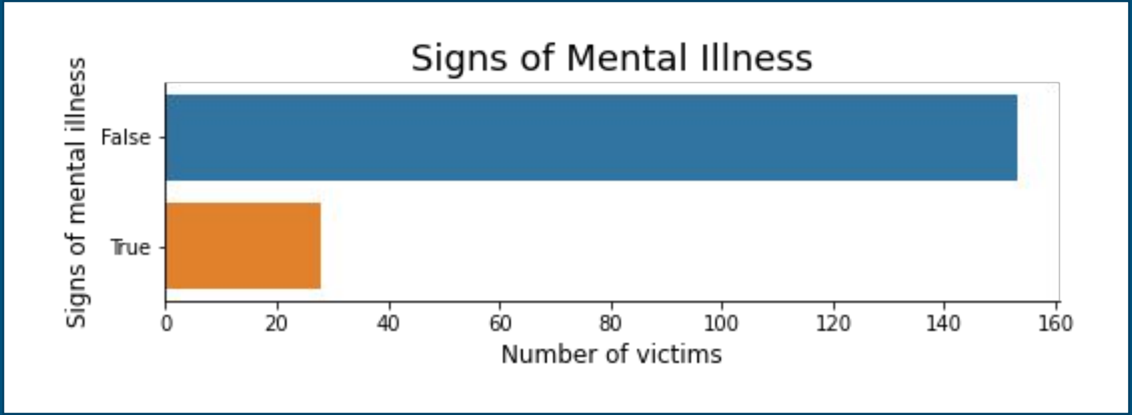
4. What was the Threat Level demonstrated by the victims?



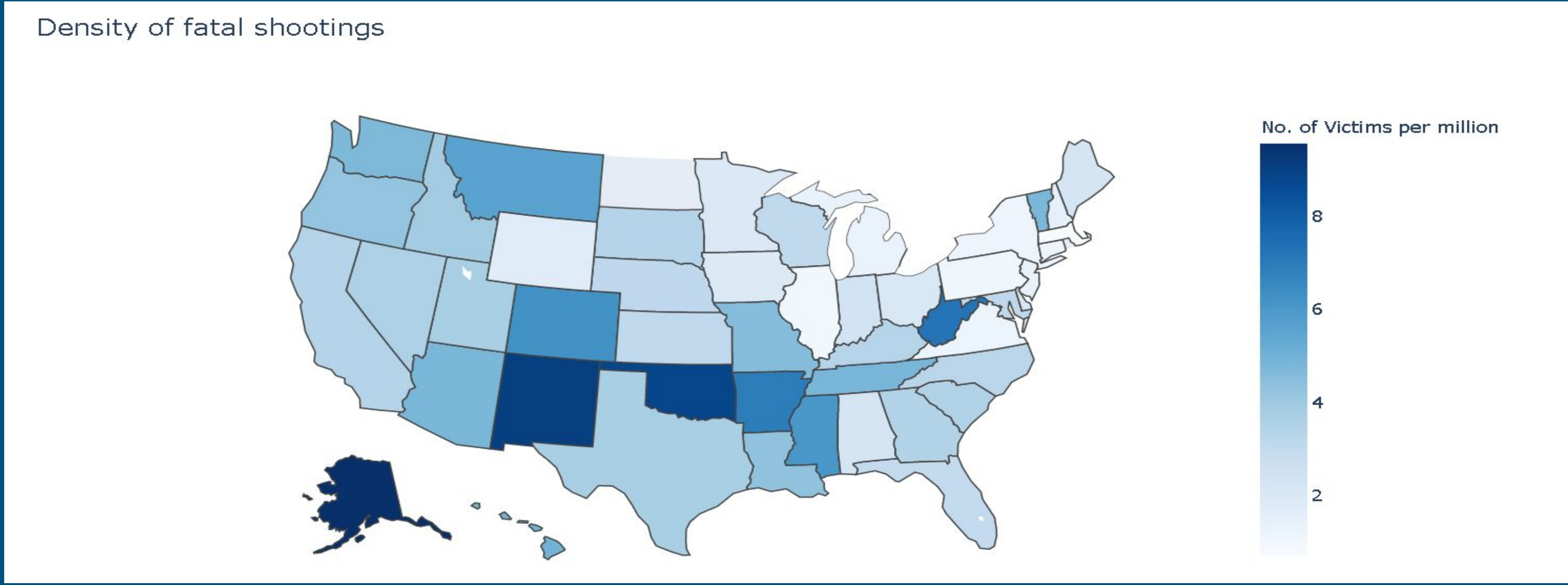
5. What was the Flee method used by the victims?



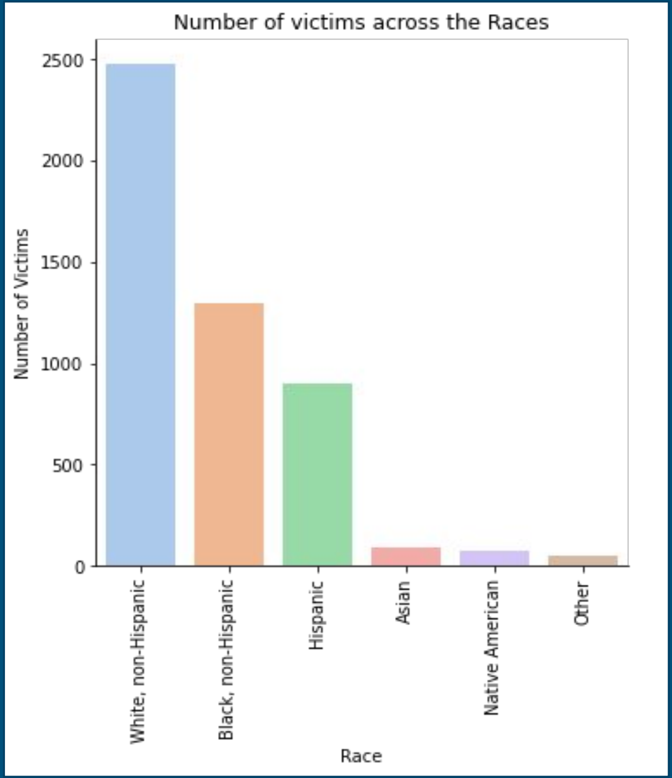
6. How many victims showed Signs of Mental Illness?

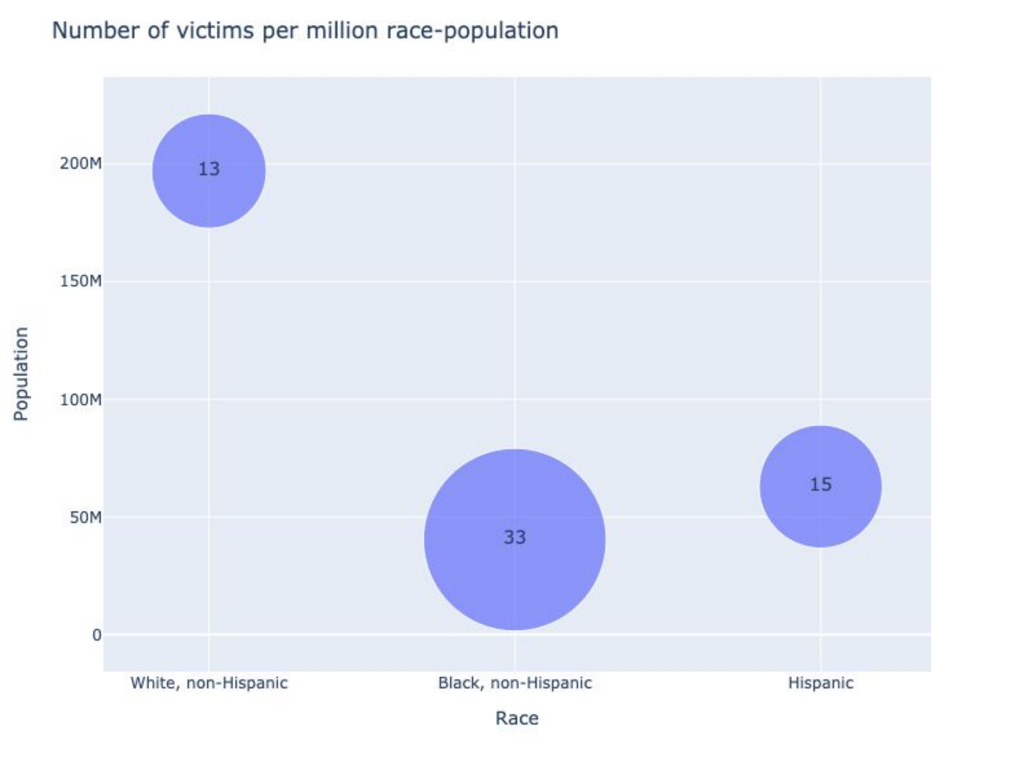


Combining with Population dataset



Ethnicity population Density





Data Cleaning

1) Null values in armed, age, gender, race, flee

- Race → imputed with ‘unknown’
- Flee, Armed*, Gender → imputed most frequent value
- Age → imputed mean

**Armed contained many categories - we kept top three categories (gun, knife, unarmed) and converted rest to ‘other’*

2) Remove irrelevant columns + create dummy variables! → final shape (5416, 19)

ML: Signs of Mental Illness

Gaussian Naive Bayes method

	precision	recall	f1-score	support
0	0.8724	0.6672	0.7561	1271
1	0.3522	0.6497	0.4568	354
accuracy			0.6634	1625
macro avg	0.6123	0.6585	0.6065	1625
weighted avg	0.7591	0.6634	0.6909	1625

Logistic Regression method

	precision	recall	f1-score	support
0	0.7863	0.9929	0.8776	1271
1	0.5500	0.0311	0.0588	354
accuracy			0.7834	1625
macro avg	0.6681	0.5120	0.4682	1625
weighted avg	0.7348	0.7834	0.6992	1625

Feature Engineering, removing Race

Gaussian Naive Bayes method

	precision	recall	f1-score	support
0	0.8310	0.7773	0.8033	1271
1	0.3509	0.4322	0.3873	354
accuracy			0.7022	1625
macro avg	0.5909	0.6048	0.5953	1625
weighted avg	0.7264	0.7022	0.7126	1625

Logistic Regression method

	precision	recall	f1-score	support
0	0.7835	0.9992	0.8783	1271
1	0.7500	0.0085	0.0168	354
accuracy			0.7834	1625
macro avg	0.7667	0.5038	0.4475	1625
weighted avg	0.7762	0.7834	0.6906	1625

ML: Race

Gaussian Naive Bayes method

labels: ['A', 'B', 'H', 'N', 'O', 'W', 'unknown'] and codes: [0, 1, 2, 3, 4, 5, 6]

	precision	recall	f1-score	support
0	0.02	0.96	0.03	27
1	0.33	0.01	0.02	383
2	0.16	0.03	0.04	280
3	0.00	0.00	0.00	18
4	0.00	0.00	0.00	12
5	1.00	0.00	0.00	759
6	0.12	0.04	0.06	146
accuracy			0.03	1625
macro avg	0.23	0.15	0.02	1625
weighted avg	0.58	0.03	0.02	1625

Logistic Regression method

labels: ['A', 'B', 'H', 'N', 'O', 'W', 'unknown'] and codes: [0, 1, 2, 3, 4, 5, 6]

	precision	recall	f1-score	support
0	0.00	0.00	0.00	27
1	0.42	0.30	0.35	383
2	0.00	0.00	0.00	280
3	0.00	0.00	0.00	18
4	0.00	0.00	0.00	12
5	0.51	0.90	0.65	759
6	0.00	0.00	0.00	146
accuracy			0.49	1625
macro avg	0.13	0.17	0.14	1625
weighted avg	0.34	0.49	0.39	1625

Feature Engineering, grouping least frequent values

Gaussian Naive Bayes method

labels: ['All Others', 'B', 'W'] and codes: [0, 1, 2]

	precision	recall	f1-score	support
0	0.37	0.25	0.30	483
1	0.38	0.36	0.37	383
2	0.53	0.65	0.59	759
accuracy			0.47	1625
macro avg	0.43	0.42	0.42	1625
weighted avg	0.45	0.47	0.45	1625

Logistic Regression method

labels: ['All Others', 'B', 'W'] and codes: [0, 1, 2]

	precision	recall	f1-score	support
0	0.47	0.16	0.23	483
1	0.45	0.28	0.35	383
2	0.52	0.84	0.64	759
accuracy			0.51	1625
macro avg	0.48	0.43	0.41	1625
weighted avg	0.49	0.51	0.45	1625

ML: Flee method

Gaussian Naive Bayes method

labels: ['Car', 'Foot', 'Not fleeing', 'Other'] and codes: [0, 1, 2, 3]

	precision	recall	f1-score	support
0	0.25	0.62	0.35	289
1	0.26	0.13	0.17	221
2	0.80	0.54	0.65	1066
3	0.03	0.04	0.03	49
accuracy			0.48	1625
macro avg	0.33	0.33	0.30	1625
weighted avg	0.60	0.48	0.51	1625

Logistic Regression method

labels: ['Car', 'Foot', 'Not fleeing', 'Other'] and codes: [0, 1, 2, 3]

	precision	recall	f1-score	support
0	1.00	0.00	0.00	289
1	1.00	0.00	0.00	221
2	0.66	1.00	0.79	1066
3	1.00	0.00	0.00	49
accuracy			0.66	1625
macro avg	0.91	0.25	0.20	1625
weighted avg	0.77	0.66	0.52	1625

Feature Engineering

Random Forest classifier method

labels: ['Car', 'Foot', 'Not fleeing', 'Other'] and codes: [0, 1, 2, 3]

	precision	recall	f1-score	support
0	0.27	0.29	0.28	289
1	0.23	0.29	0.26	221
2	0.75	0.61	0.67	1066
3	0.01	0.04	0.02	49
accuracy			0.50	1625
macro avg	0.31	0.31	0.31	1625
weighted avg	0.57	0.50	0.53	1625

Random Forest classifier method with reduced features

labels: ['Car', 'Foot', 'Not fleeing', 'Other'] and codes: [0, 1, 2, 3]

	precision	recall	f1-score	support
0	0.22	0.35	0.27	289
1	0.15	0.24	0.18	221
2	0.76	0.34	0.47	1066
3	0.02	0.16	0.04	49
accuracy			0.32	1625
macro avg	0.29	0.27	0.24	1625
weighted avg	0.56	0.32	0.38	1625

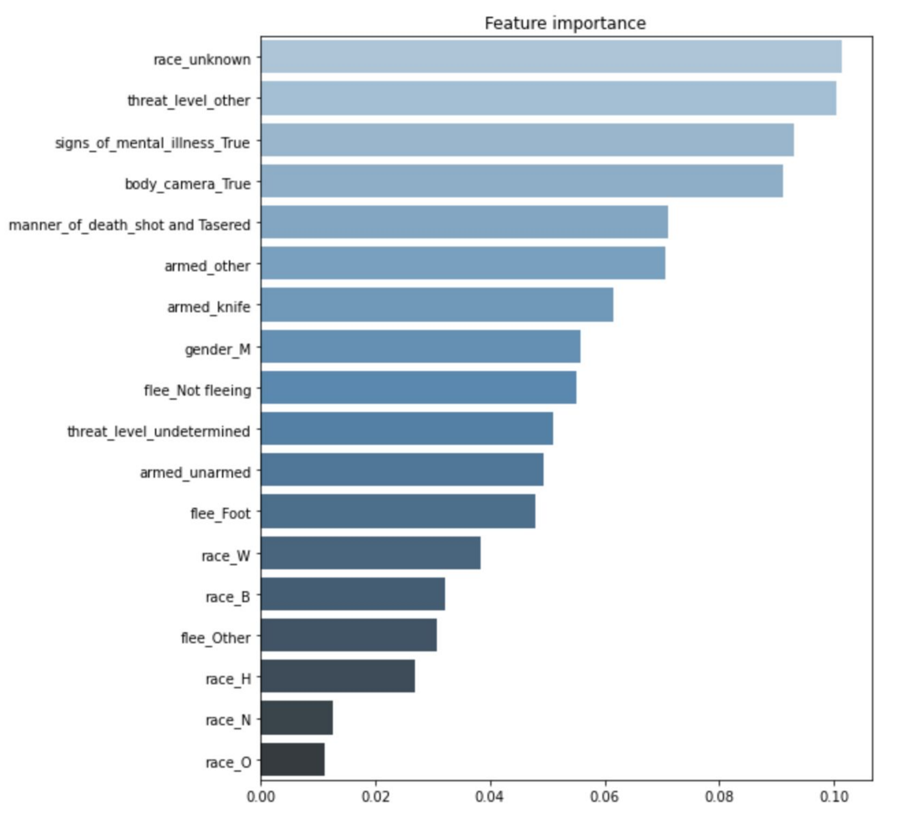
ML: Age

Linear regression model

```
Coefficients:[ 0.25575037 -1.84920101 -1.72157503
-4.23929447  1.02696981 -2.94600306
 -1.85789835 -3.77667446 -2.7277103   3.46849395
6.15051486  1.09913004
 -0.4055692  -0.86735236 -1.25676577  4.28709909
-1.25611337 -0.38608253],
intercept 33.08
Residual sum of squares: 142.58
R-squared Score: 0.13
```

Remove dummy variables of race

```
Coefficients:[ 0.12828654 -2.02845854 -1.83161315
-5.30711232  0.59033374  1.9860364
 -0.57752548 -0.60086937 -2.16621504  4.41043883
-1.50463908 -1.22628728],
intercept 34.67
Residual sum of squares: 153.62
R-squared Score: 0.07
```



Conclusion

- California has the highest number of deaths so we set it as a base for our storyline and presented metrics related to it.
- Population Density:
  - Alaska has the highest density of deaths across states.
  - Whites have the highest number of deaths in number. However, in proportion, the Black victims per million rate is more than twice the Whites' .
- Only model that was accurate enough was about mental Illness.

Restrictions and limitations faced

Limitation	Details and solution
Data	● Our dataset had largely categorical data which had its limitations in terms of generating visualisations
ML	● Only one model gave decent results ● Most of the variables were imbalanced which led to poor performing models

Thank you for your attention!

Any Questions?