Fatal Force

The dataset I have chosen to study contains records of every fatal shooting in the United States by a police officer in the line of duty since Jan 1, 2015. The Data was obtained from **The Washington Post's** database which contains data about each fatal shooting in CSV format which could be easily downloaded from github.

The dataset includes observations for specific details of the individual such as race, longitude, latitude and mental-health signs, body camera, threat level and arms - using information procured from local news reports, law enforcement websites and social media. It has been brought to light that the fatal shootings reported in this dataset by **The Post** are twice as many as those recorded by the FBI on average annually and it is updated regularly as facts emerge about individual cases. **So it is very crucial for us to know the plight of a particular community in the US with correct numbers unlike the analysis drawn from the news that comes straight away from the authorities and that is something we can achieve using this dataset.**

Import the Dataset

The data, as we can see, poses some challenges to our insights and it is very important to make it clear. First, we don't have regular "id" throughout the data and this is in addition to the missing observations for few variables which we will see as we proceed.
The dataset has 6728 observations and 17 variables originally and this could be verified using the following codes and their outputs.
The missing values and data types for variables as identified are:

```
id                          0
name                      270
date                        0
manner_of_death             0
armed                     206
age                       324
gender                      5
race                      875
city                        0
state                       0
signs_of_mental_illness     0
threat_level                0
flee                      519
body_camera                 0
longitude                 318
latitude                  318
is_geocoding_exact          0
dtype: int64
```

```
id                          int64
name                       object
date                       object
manner_of_death            object
armed                      object
age                       float64
gender                     object
race                       object
city                       object
state                      object
signs_of_mental_illness      bool
threat_level               object
flee                       object
body_camera                  bool
longitude                 float64
latitude                  float64
is_geocoding_exact           bool
dtype: object
```

EDA

The 5 states that saw least shootings in the past 7 years and those with highest shootings for the same period of time are given as arranged in ascending order,

Head

state	count
RI	4
VT	11
WY	15
ND	15
DE	16

Tail

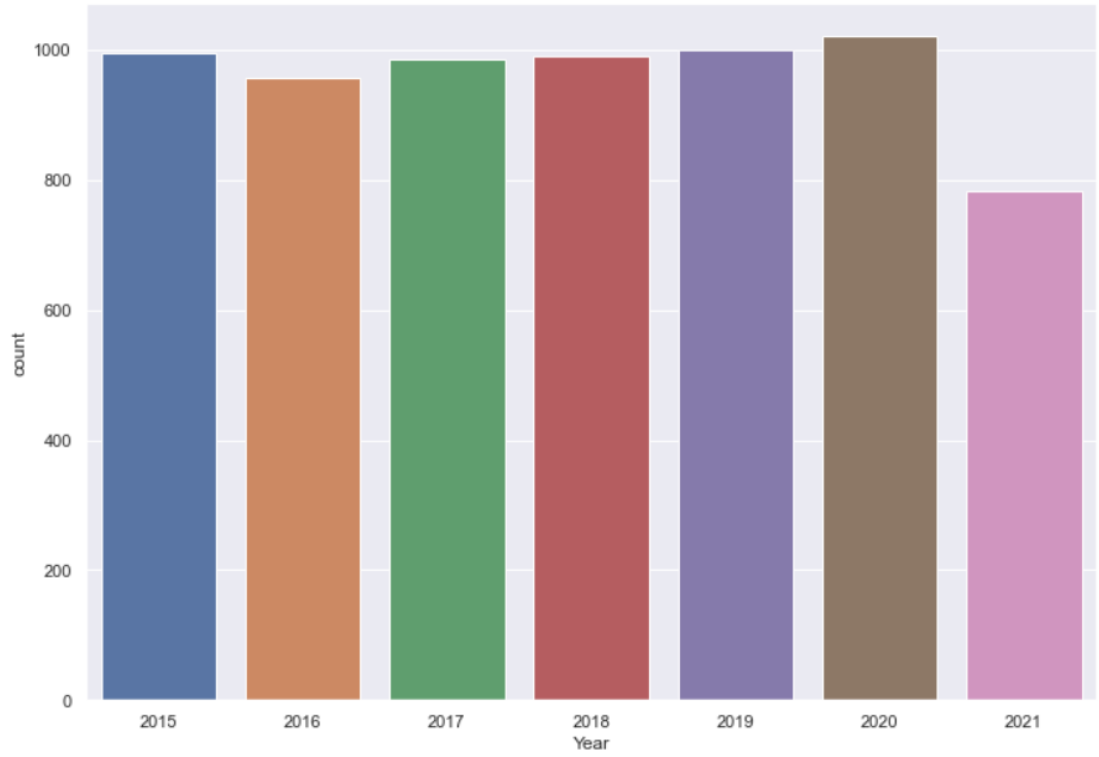
state	count
GA	246
AZ	305
FL	437
TX	601
CA	988

The exact state names can be looked into any government directory but if any Python packages offer to help, I would be most interested to know them.
Since [“age”, “geography”, “id”] were the only numeric variables in the dataset, the point of describing them made very little sense and thus proceeded with the objects and found some significant results.

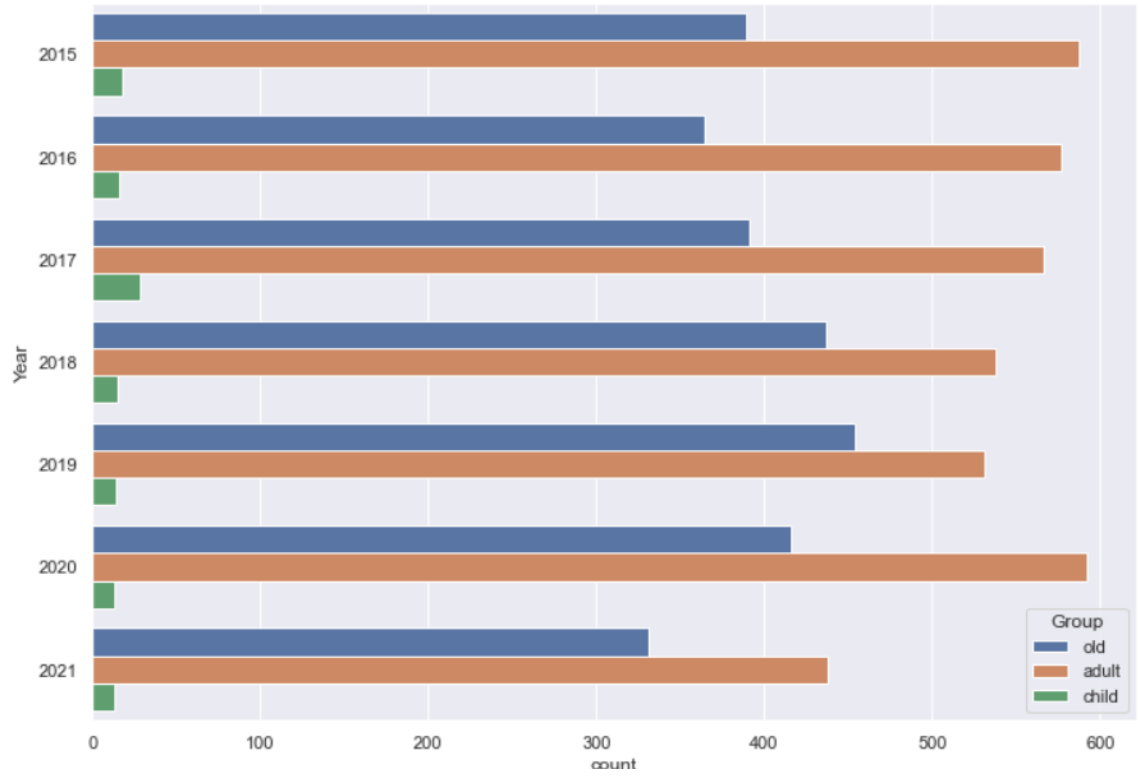
	name	manner_of_death	armed	gender	race	city	state	threat_level	flee
count	6458	6728	6522	6723	5853	6728	6728	6728	6209
unique	6437	2	98	2	6	2863	51	3	4
top	Michael Johnson	shot	gun	M	W	Los Angeles	CA	attack	Not fleeing
freq	3	6395	3872	6423	2969	105	988	4333	4007

So, the missing values for race could be attributed to the individual’s right about voluntary disclosures.
I then decided to group people and my expectation was that children would not be there.

Plot using Seaborn

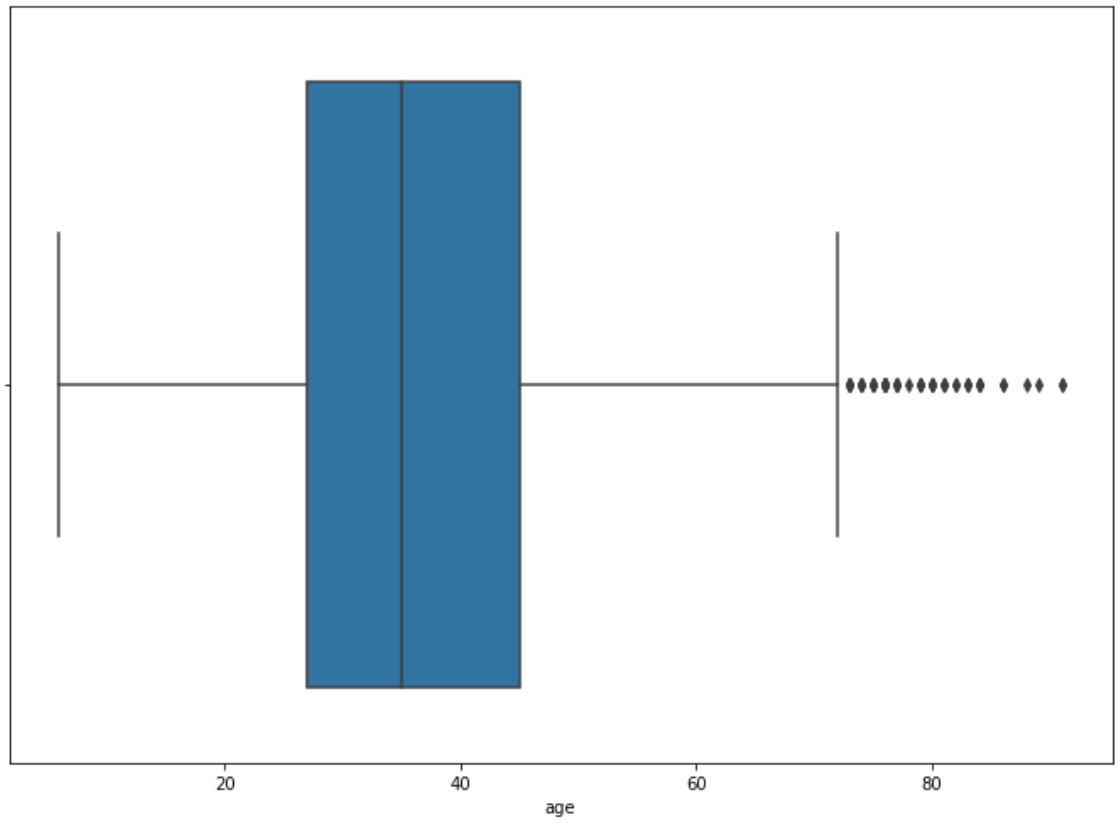


As I dug deeper, we found that the number of shootings in The Post's data, which identified a lot more fatal forces, has been closer to each other every year since 2015 except for 2021.



I was proved wrong and found out that children (number being less than 100) were also the victims of fatal police shootings.

Box plot for Age

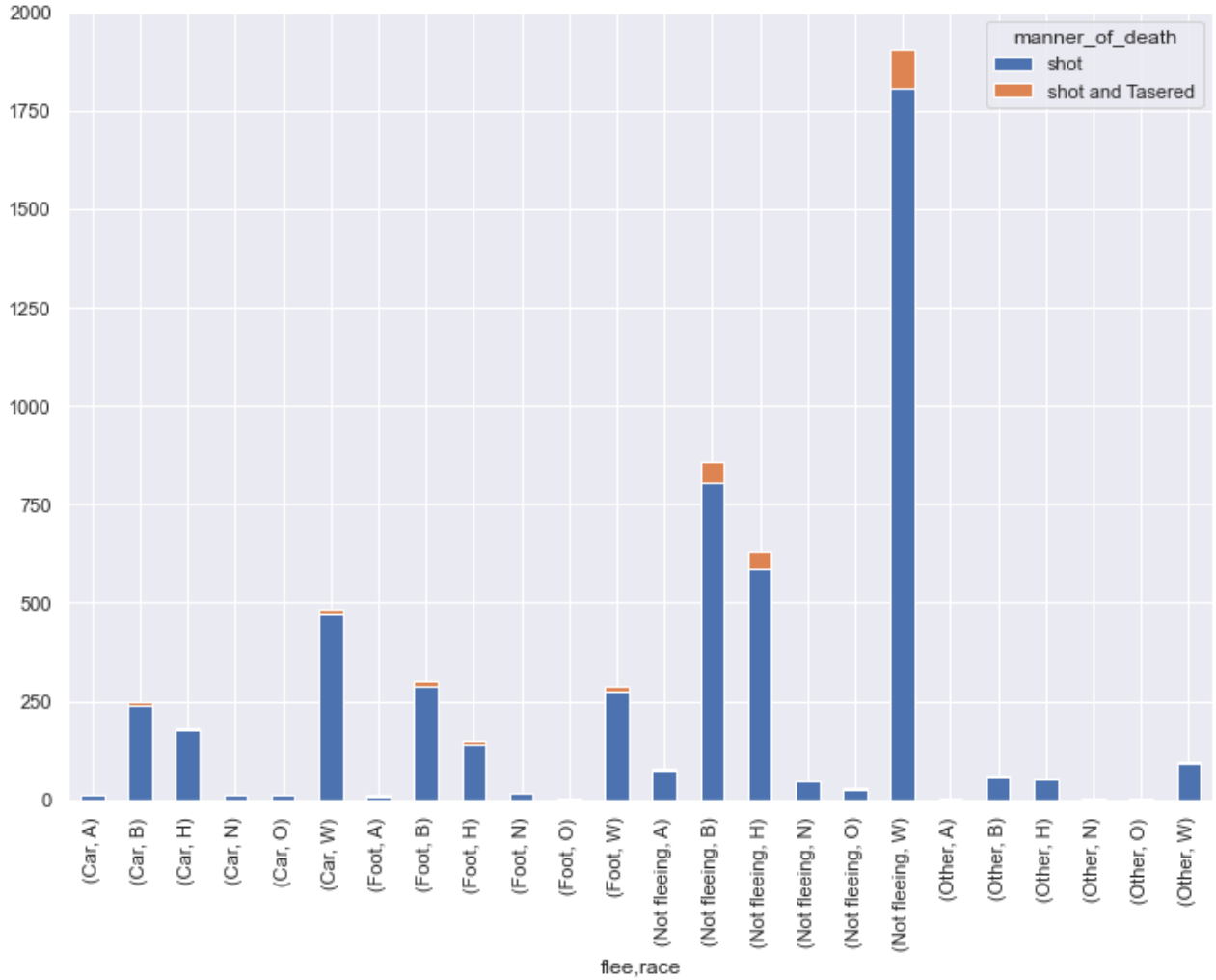


Contingency Tables

manner_of_death			shot	shot and Tasered
body_camera	signs_of_mental_illness	threat_level		
False	False	attack	2889	93
		other	1305	84
		undetermined	161	2
	True	attack	750	43
		other	400	45
		undetermined	24	0
True	False	attack	391	20
		other	225	21
		undetermined	20	2
	True	attack	136	11
		other	87	12
		undetermined	7	0

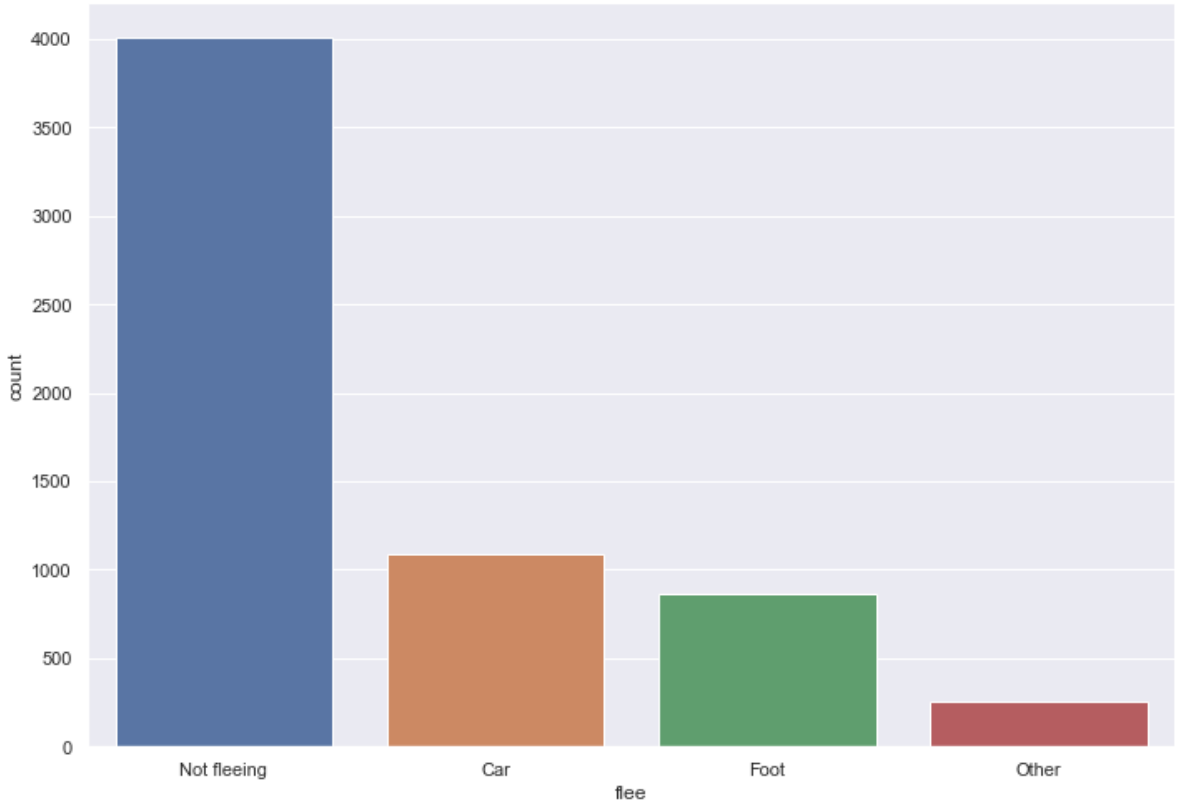
manner_of_death		shot	shot and Tasered
flee	threat_level		
Car	attack	690	13
	other	344	13
	undetermined	29	0
Foot	attack	537	27
	other	261	17
	undetermined	18	1
Not fleeing	attack	2527	104
	other	1168	119
	undetermined	87	2
Other	attack	151	11
	other	76	4
	undetermined	9	1

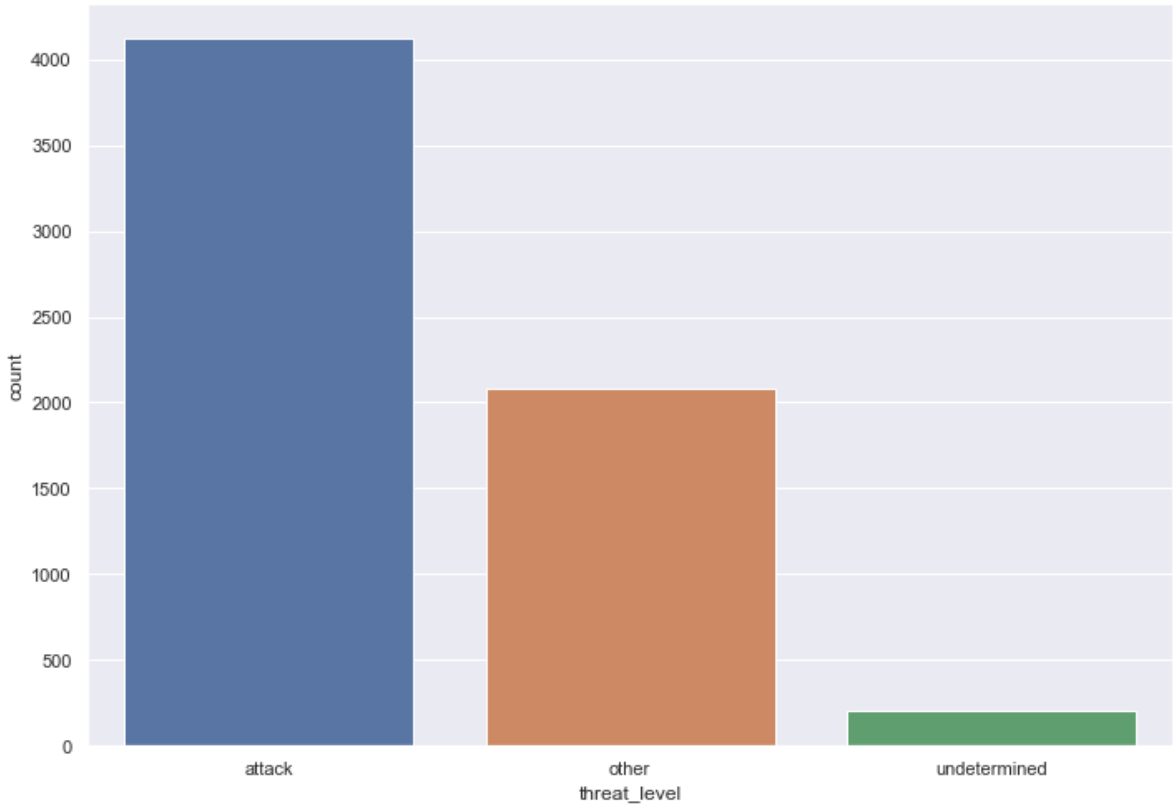
Not fleeing but attacking looks like a big problem for the people which might lead to confusion which further becomes the root cause of the mishap.



Clearly, Not_fleeing and W race have majorly faced police shootings.

Evidences from crosstabs





Therefore, my limited knowledge suggests that, to draw insights from this dataset I would be more likely to use classification models as numerical values in my data are of no significance or a better way to put it “very less significant”.

Missing values analysis

```
gun                     3679
knife                    931
unarmed                  410
toy weapon               226
vehicle                  205
undetermined             184
unknown weapon            84
machete                   50
Taser                     34
sword                     25
ax                        24
gun and knife             22
hammer                    19
baseball bat              19
screwdriver               17
gun and vehicle           16
sharp object              15
metal pipe                15
hatchet                   14
BB gun                    14
box cutter                13
gun and car               11
scissors                   9
piece of wood              8
crossbow                   8
vehicle and gun            8
shovel                     7
pipe                       7
rock                       6
straight edge razor        5
Name: armed, dtype: int64
```

After imputation of important variables with mode:

```
armed                   0
race                    0
body_camera             0
signs_of_mental_illness 0
flee                    0
threat_level            0
Group                   0
dtype: int64
```

Since the variables with missing values are categorical, I applied a common solution to such imputations. The missing values were filled with mode or most common observation for the specific variable.

Classification

The problem I am trying to solve is the identification of threats with the given details of individuals. Thus, my target variable is threat level to be predicted with a classification model with the remaining categorical variables.

The train-test split data accounted for 30:70 of the original data.

The curse of dimensionality has also been taken into consideration and PCA solved the problem to some extent as there were a lot of new columns added because of the new dummy variables created for each column in the original data.

I tried running logistic regression but it failed as the data frame obtained after creating dummy variables returned a **singular matrix.**

Moving further I chose KNN classifier and it returned great results with test accuracy of more than 97% making it a great classifier for the data.

But, it is important to understand that KNN converts the values into numeric and the indices are changed too. So, they should not be confused with those of original dataframe,