DATA SCIENCE -351

PROJECT

VINIT, SERGEI, MATTHEW

PROJECT #3: FATAL FORCE IN THE US

CONTENT

DATA CLEANING

The process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

DATA CLEANING

Initial Data

id	2028
name	2028
date	2028
manner_of_death	2028
armed	2022
age	1991
gender	2028
race	1937
city	2028
state	2028
signs_of_mental_illness	2028
threat_level	2028
flee	2001
body_camera	2028

Training Data

- Imported Files include training and testing data. Along with optional share race by city
- Observed number of rows is 2028
- Armed, Age, Race and Flee has missing rows to be filled in

- Filling the missing age we use median value, 34 years, to fill in missing rows.
- For Armed rows we delete 6 rows having a large sample.
- For Race and Flee type we use probability of each caterogy occurace and randomly fill in the data

Cleaned data

id	2022
name	2022
date	2022
manner_of_death	2022
armed	2022
age	2022
gender	2022
race	2022
city	2022
state	2022
signs_of_mental_illness	2022
threat_level	2022
flee	2022
body_camera	2022

DATA CLEANING CONTINUED...

Testing data

- Applying the same steps for the testing data
- Final number of rows reduced from 507 to 504
- We observe that additional data set, share of race per city, doesnt have any missing data

Initial Data

```
id                         507
name                       507
date                       507
manner_of_death            507
armed                      504
age                        467
gender                     507
race                       403
city                       507
state                      507
signs_of_mental_illness    507
threat_level               507
flee                       469
body_camera                507
dtype: int64
```

Cleaned data

```
id                         504
name                       504
date                       504
manner_of_death            504
armed                      504
age                        504
gender                     504
race                       504
city                       504
state                      504
signs_of_mental_illness    504
threat_level               504
flee                       504
body_camera                504
```

EXPLORATORY DATA ANALYSIS (EDA)

Initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

01

QUESTION 1

Which state has the most fatal police shootings? Which city is the most dangerous?

02

QUESTION 1

What is the most common way of being armed?

03

QUESTION 1

What is the age distribution of the victims? Compare age distribution of different races?

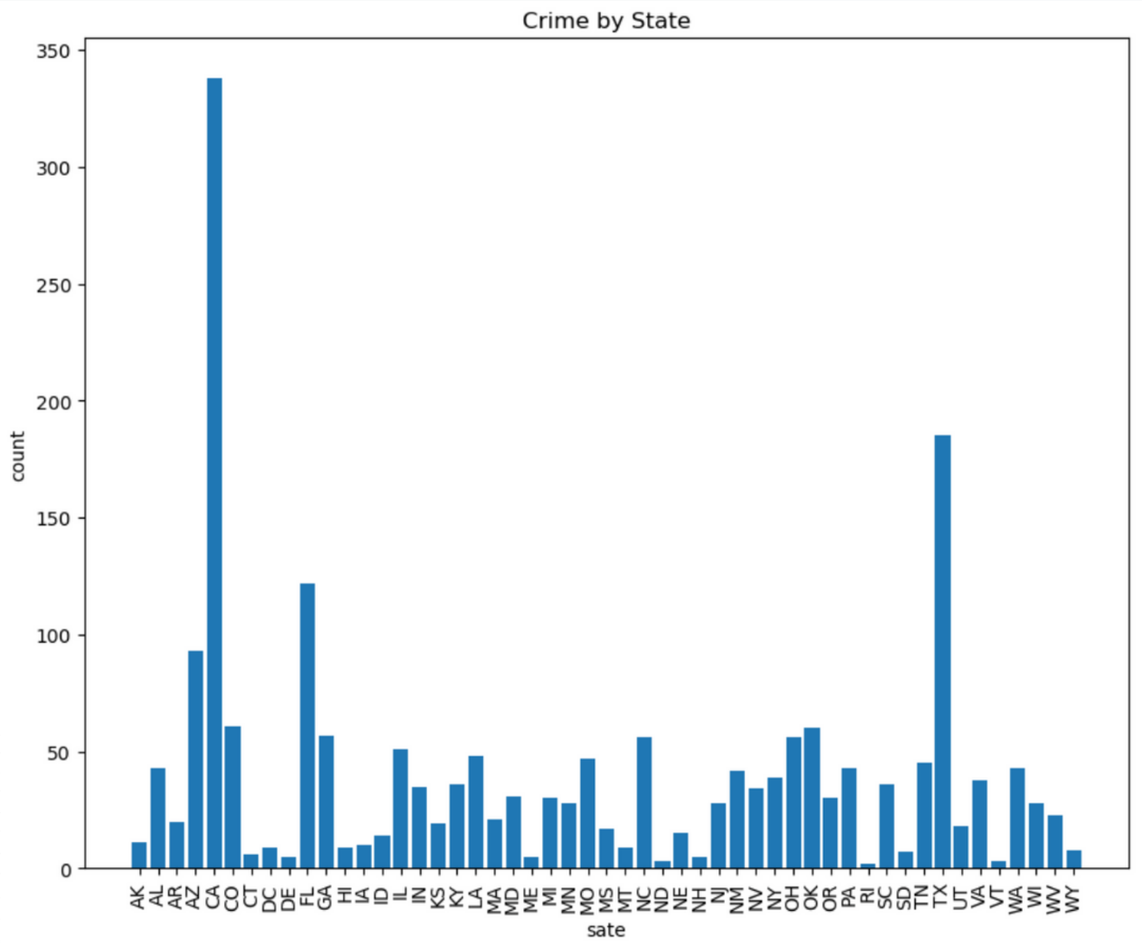
04

QUESTION 4

Compare the total number of people killed per race. Compare the number of people killed per race as a proportion of respective races. What difference do you observe?

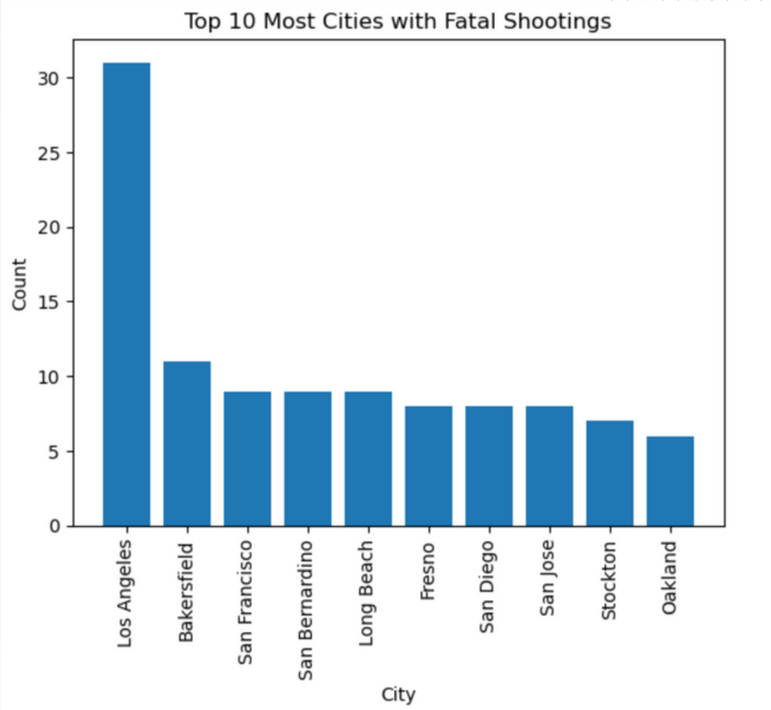
01

- **Which state has the most fatal police shootings? Which city is the most dangerous?**



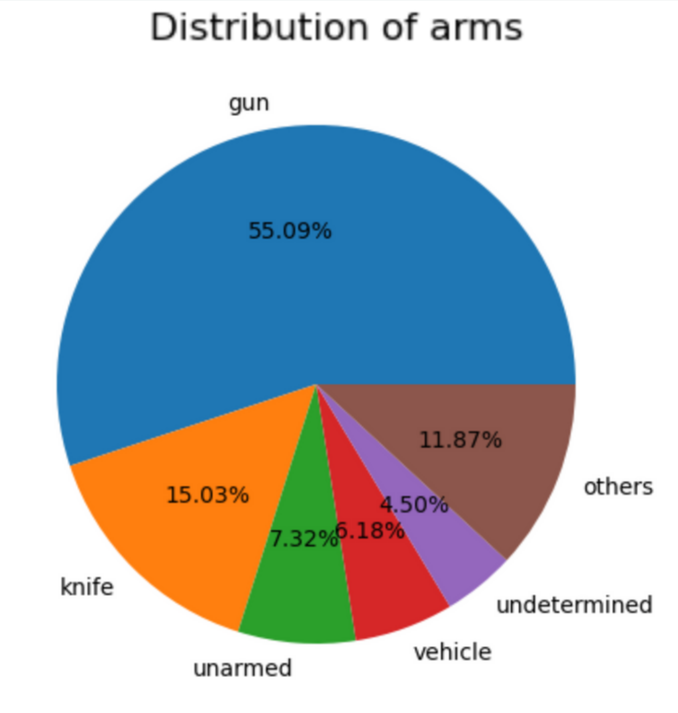
- California is the state with most fatal police shootings with a total 338 recorded incident, followed by TX with 186 of cases California has.

- Los Angeles is the most dangerous city in terms of police killings with 31 such cases in the city alone. Followed by Bakersfield with 11 cases recorded.



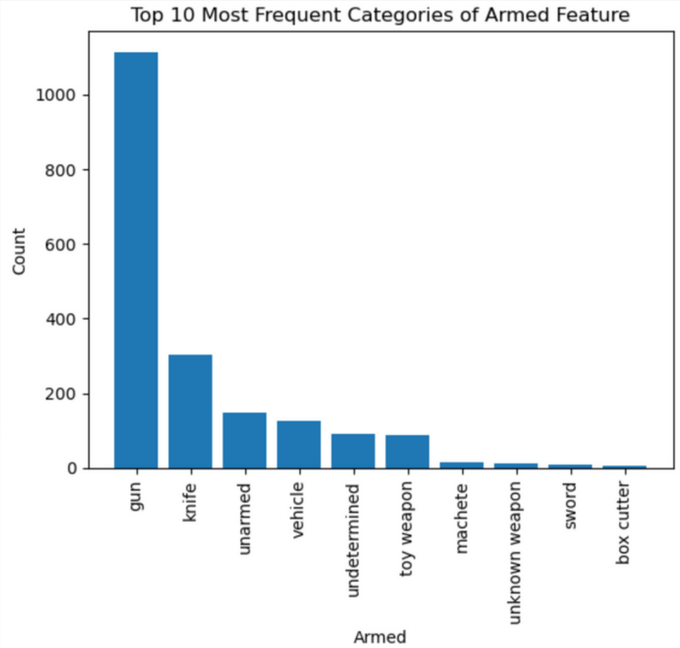
02

- **What is the most common way of being armed?**



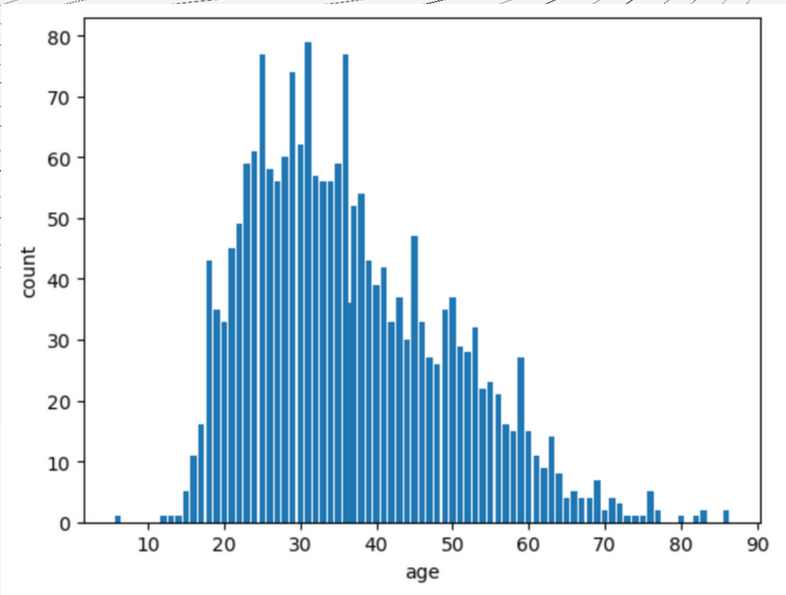
- The most common way of being armed is carrying a gun with 1114 recorder incidents being 55.09%,

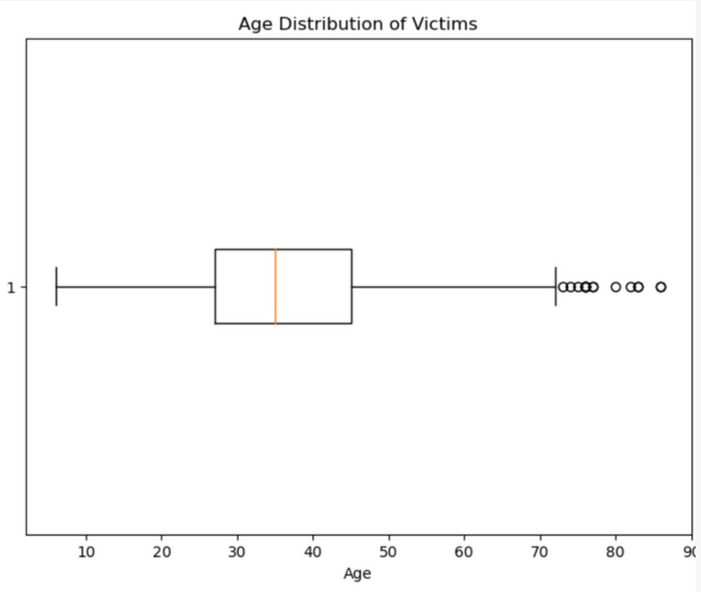
- Second common is by carrying a knife, 304, with 15.03%



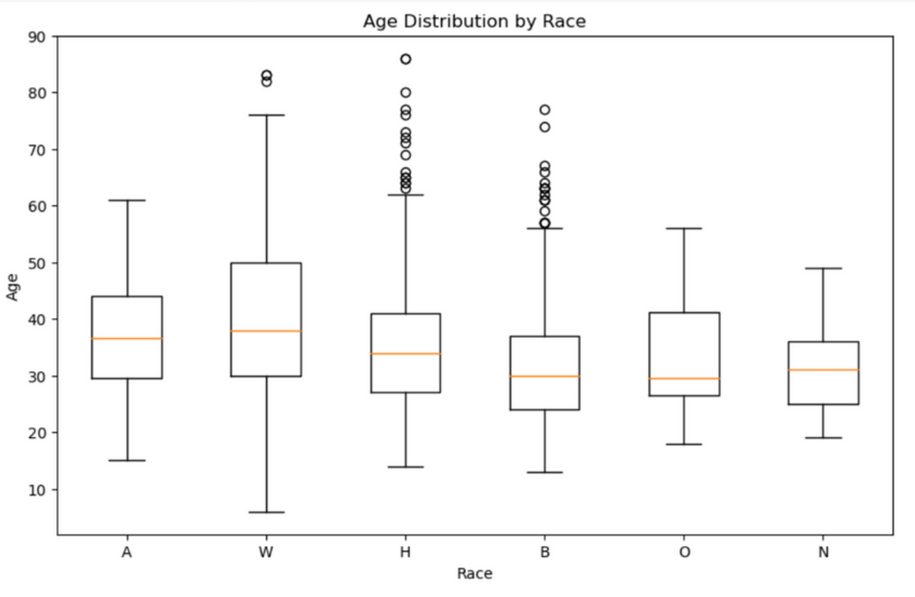
03

- **What is the most common way of being armed?**





- The Median age of victims killed due to police killings is 35, therefore 50%.
- Individuals age 27-45 are most likely to be victims of police killings

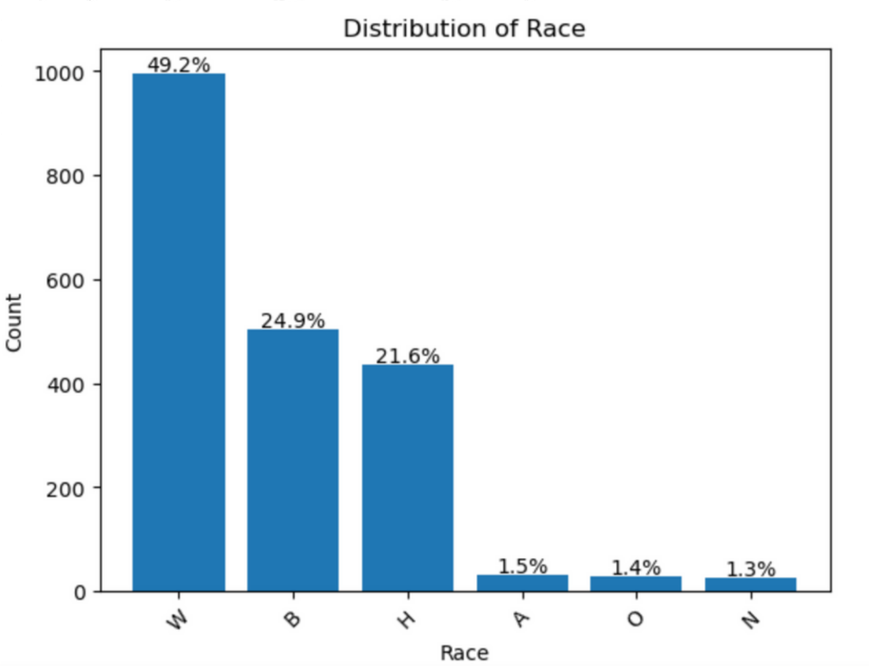


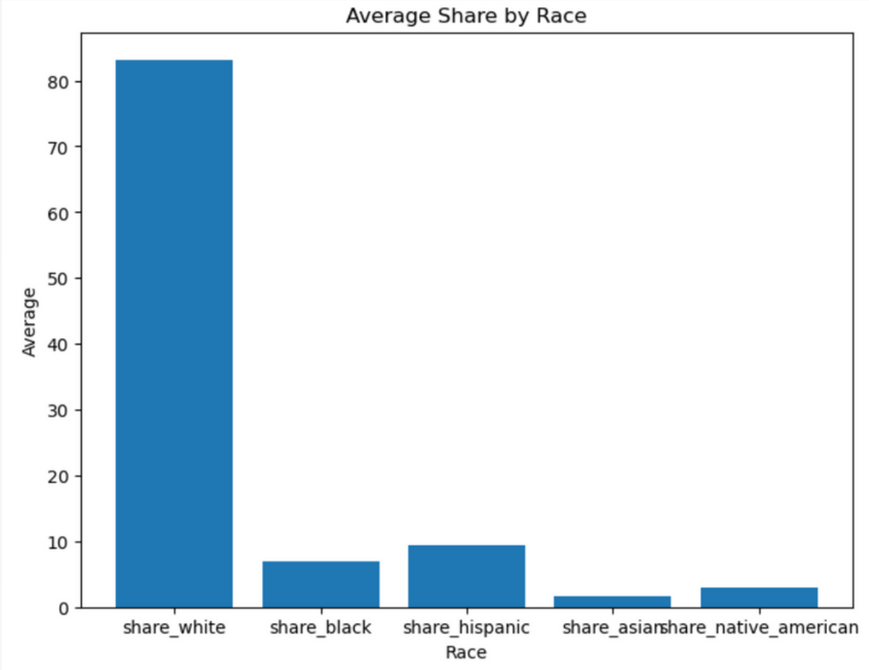
- Across all races, the age range of most of the victims was late twenties to early forties.
- The youngest person killed was found to be 6 years old
- The oldest person killed was 86 year old White person.
- The most frequently occurring age in our data set was found to be 31 years.

count	2022.000000
mean	36.565233
std	12.776698
min	6.000000
25%	27.000000
50%	35.000000
75%	45.000000
max	86.000000

04

- **Compare the total number of people killed per race. Compare the number of people killed per race as a proportion of respective races. What difference do you observe?**



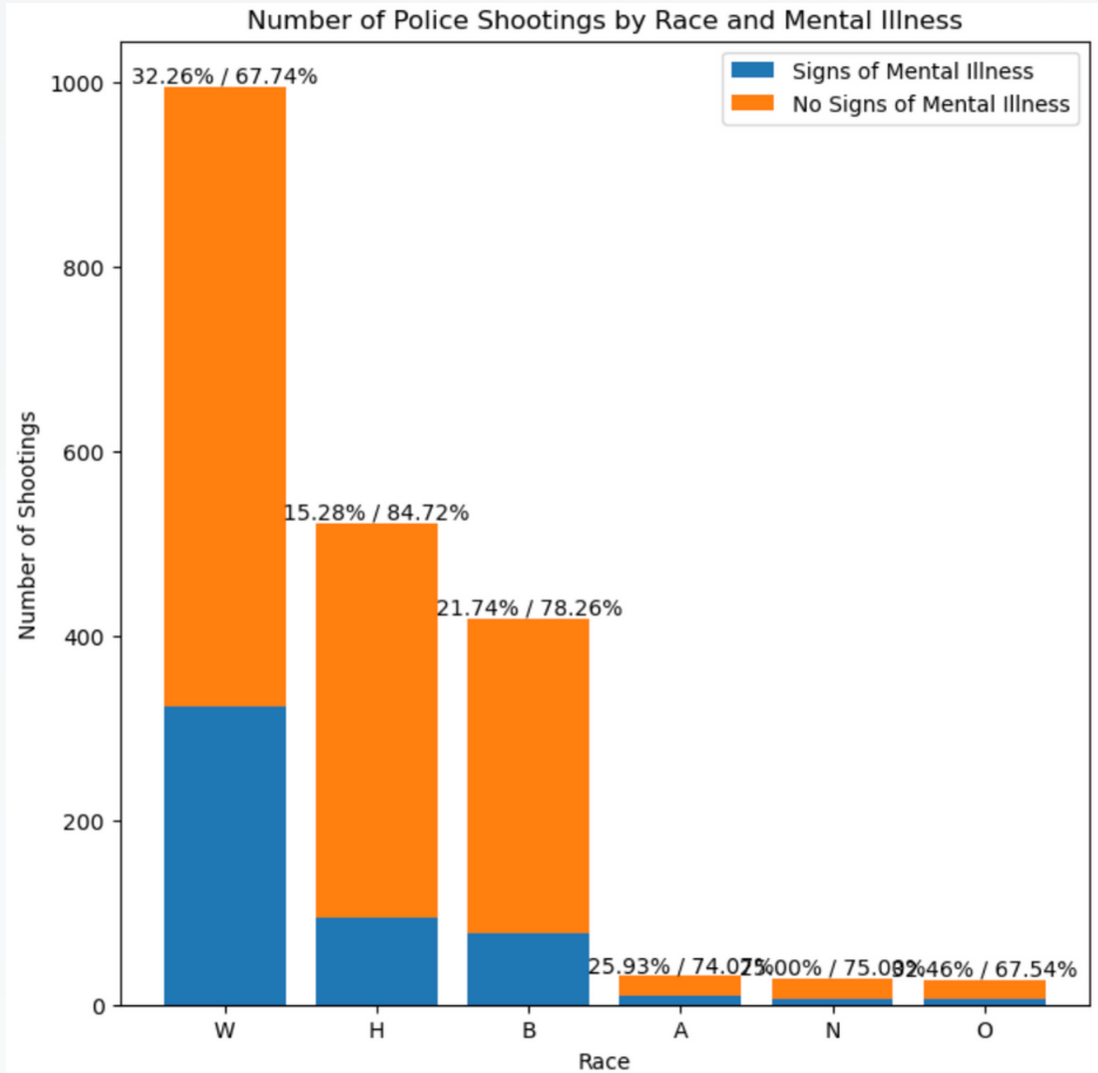


- out of all the shooting about 50% involved race 'white' but comparing to the ration of the population we notice that 'white' population consist of more than 80% of the population. However, we notice that the next most fatal shooting incidents involve 'black' race having 25% of shooting incidents, but their share ration across population is only about 9%, this signifies that more fatal shooting by police occur with 'black' race comparing their population size. Additionally we notice that 'hispanic' race share 22% of the shooting incidents and their share of the population is about 10%.

EXTRA EDA QUESTION...

QUESTION 1

- **How was the distribution of the victims identified with a mental illness between the different races?**



Based on the results found we see that the biggest ration of metal illness and no metal illness is shared across **'white'** race with ration of 33/67, meaning that the shooting could be justified by the state of the victim.

Data Preprocessing

```
race_map = {"W":1,"B":2,"A":3,"N":4,"H":6,"O":5, 1:1, 2:2, 3:3, 4:4, 5:5, 6:6}

gender_map = {"M":1,"F":2,1:1,2:2}

flee_map = {"Not fleeing":1,"Car":2,"Foot":3,"Other":4,1:1,2:2,3:3,4:4}

threat_level_map = {"attack":1,"other":2,"undetermined":3,1:1,2:2,3:3}

death_map = {"shot":1,"shot and Tasered":2,1:1,2:2}

mental_illness_map = {True:1,False:2,1:1,2:2}

body_camera_map = {True:1,False:2,1:1,2:2}
```

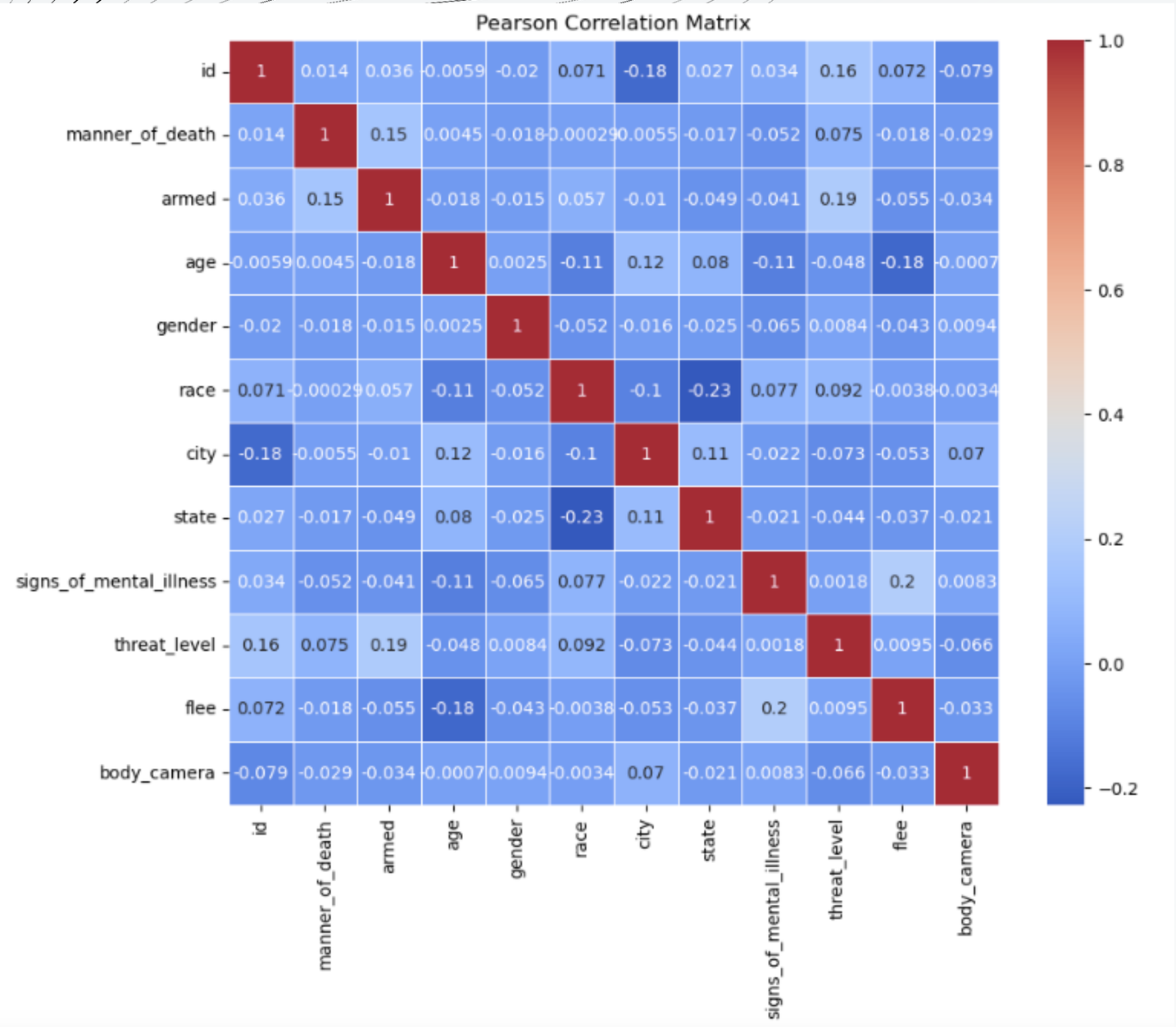
The training and testing data needed to be converted Into numerical values, starting at 1, so that it could be used by the machine learning algorithms.

Armed, City, and State were converted Into numerical data by counting the unique values and giving them a number.

NEW FORMAT OF THE DATA

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level	flee	body_camera
0	3	Tim Elliot	02/01/15	1	1	53.0	1	3	471	14	1	1	1	2
1	4	Lewis Lee Lembke	02/01/15	1	1	47.0	1	1	156	25	2	1	1	2
2	5	John Paul Quintero	03/01/15	2	3	23.0	1	6	127	33	2	2	1	2
3	8	Matthew Hoffman	04/01/15	1	6	32.0	1	1	21	1	1	1	1	2
4	9	Michael Rodriguez	04/01/15	1	50	39.0	1	6	301	5	2	1	1	2

Data Preprocessing cont.



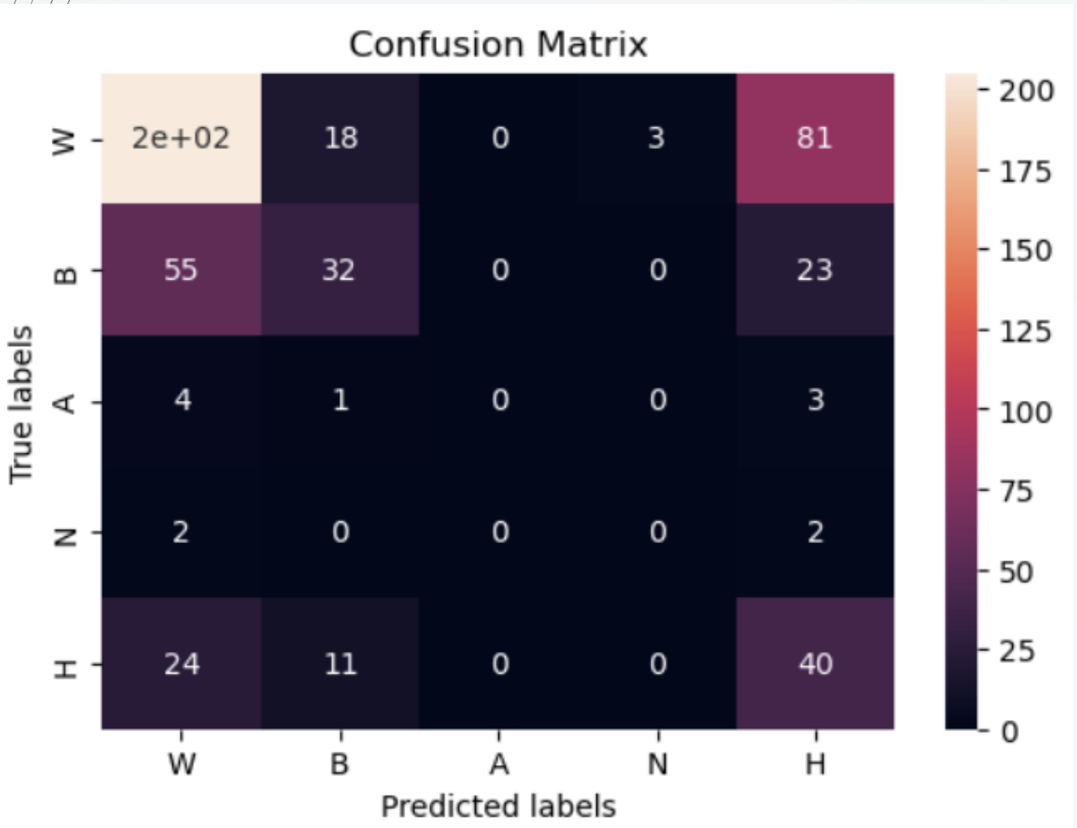
The Pearson correlation of the training data showed small correlations.

PREDICTION AND MODELLING

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate nulla at ante rhoncus, vel efficitur felis condimentum. Proin odio odio.

1 Classification using Multinomial Logistic Regression

It's a classification method that generalizes logistic regression to multiclass problems. Logistic Regression is usually binomial, therefore different functions are used for multinomial outcomes.

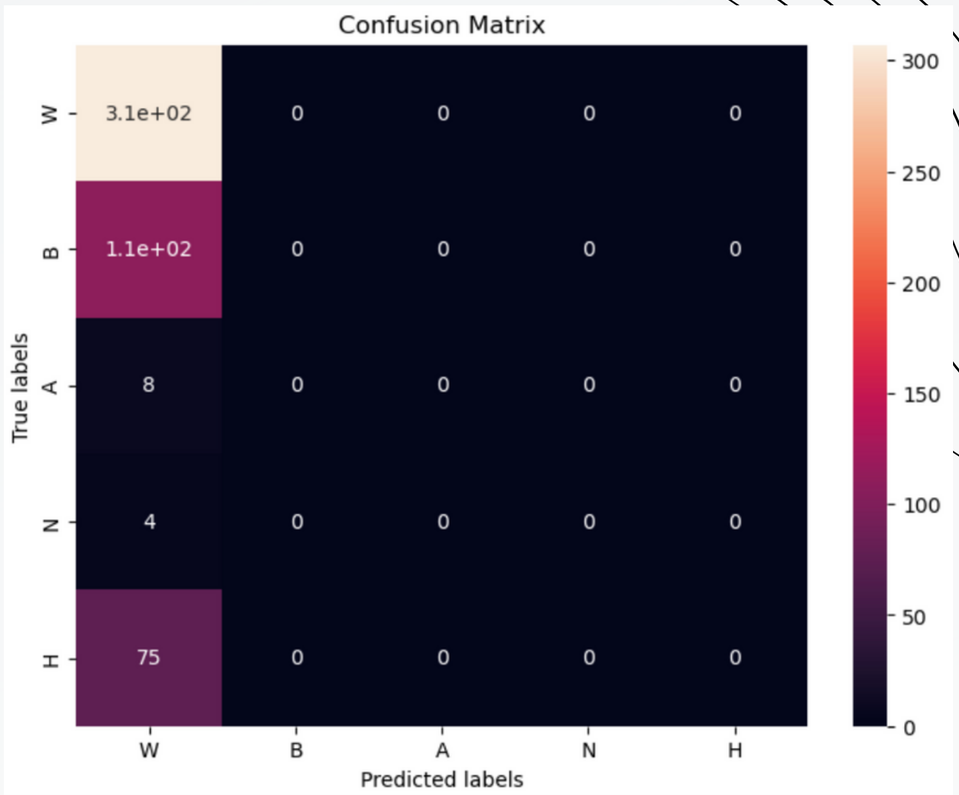


	precision	recall	f1-score	support
1	0.69	0.70	0.70	307
2	0.35	0.42	0.38	110
3	0.00	0.00	0.00	8
4	0.00	0.00	0.00	4
6	0.30	0.24	0.26	75
accuracy			0.56	504
macro avg	0.27	0.27	0.27	504
weighted avg	0.54	0.56	0.55	504

Overall Accuracy of the model is : 55.55555555555556 %

2. CLASSIFICATION USING K NEAREST NEIGHBOURS CLASSIFIER

- Type of supervised learning method used for classification.
- It determines the classification of a new data point based on the classifications of its K nearest neighbours.
- The optimal value for K is determined as the square root of the total number of samples in the training data, which is 2022. Therefore, **the value of K is set to 45**.



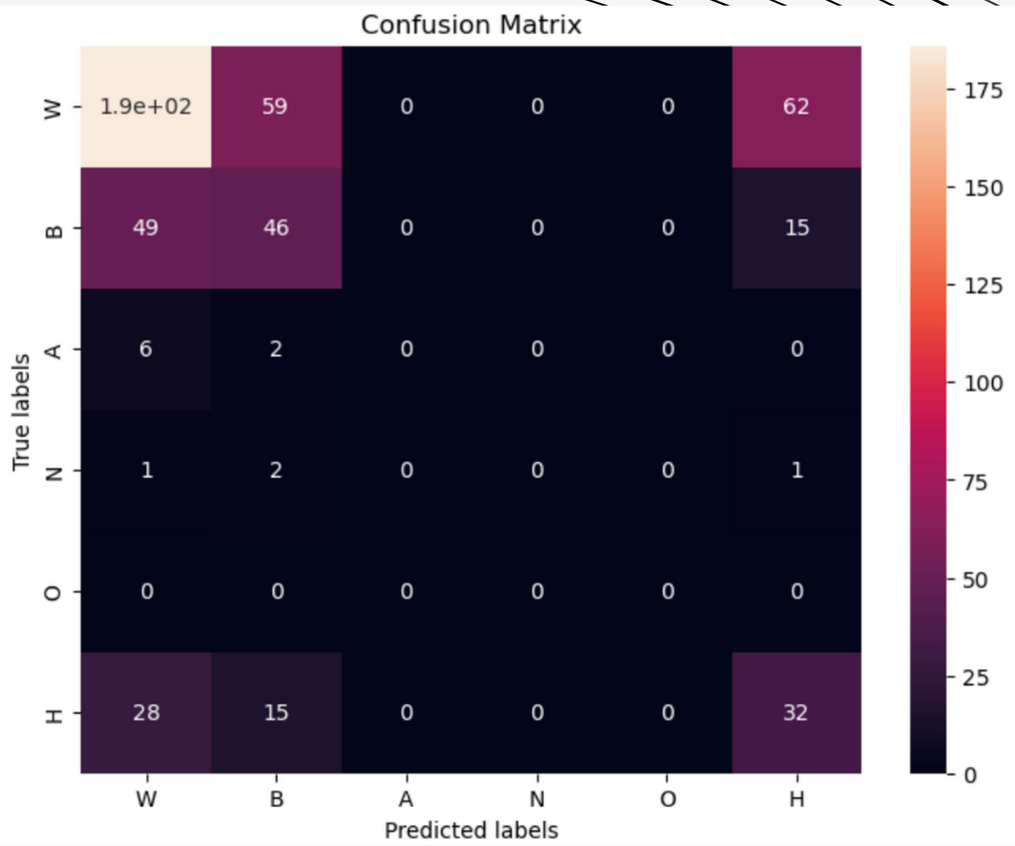
	precision	recall	f1-score	support
1	0.61	1.00	0.76	307
2	0.00	0.00	0.00	110
3	0.00	0.00	0.00	8
4	0.00	0.00	0.00	4
6	0.00	0.00	0.00	75
accuracy			0.61	504
macro avg	0.12	0.20	0.15	504
weighted avg	0.37	0.61	0.46	504

Overall Accuracy of the model is : 60.912698412698404 %

	W	B	A	N	H
W	307	0	0	0	0
B	110	0	0	0	0
A	8	0	0	0	0
N	4	0	0	0	0
H	75	0	0	0	0

3. CLASSIFICATION USING DECISION TREE CLASSIFIER

- A supervised learning technique that utilizes a structure resembling a flowchart.
- Each internal node of the tree represents a test performed on a specific attribute.
- The leaf nodes, on the other hand, represent class labels.
- The tree continues to split into nodes until either the maximum depth is reached or the purity threshold for each leaf node is achieved.
- Max depth is found by adding 1 to the number of features.
- Thus, max_depth value of decision tree is **4**



	precision	recall	f1-score	support
1	0.69	0.61	0.64	307
2	0.37	0.42	0.39	110
3	0.00	0.00	0.00	8
4	0.00	0.00	0.00	4
6	0.29	0.43	0.35	75
accuracy			0.52	504
macro avg	0.27	0.29	0.28	504
weighted avg	0.54	0.52	0.53	504

Overall Accuracy of the model is : 52.38095238095239 %

	W	B	A	N	H
W	307	0	0	0	0
B	110	0	0	0	0
A	8	0	0	0	0
N	4	0	0	0	0
H	75	0	0	0	0

CONCLUSION

- We found that Logistic Regression gave an overall accuracy of 55%.
- The Decision Tree Clasifier gave an overall accuracy of about 54%.
- The K Nearest Neighbour classifier with k = 45.

- We saw that the pearson correlation with target variable race was low.
- Our features were not the best predictor for race.
- While cleansing the data sets we found that there were a high number of missing values in our dataset in fields 'age', 'race', 'flee' and a few in 'armed' field.
- Even though we imputed the missing values, the presence of a high number of these values might have led to the low accuracy of the model.

THANK YOU!