**Exploratory Data Analysis Report on the US Shootings Data set**

By

Joshua Sebu, Akua Gyanba Morgan Acquah & Gabriel Assamah

1st June 2021

**Exploratory Data Analysis (EDA)**, also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the data set being used. To understand the US shootings data set better, we performed activities of extracting important variables and leaving behind the not too necessary variables. We tried to identify outliers, missing values, or human error that may be in the data set. We tried to also understand the relationship(s), or lack of relationship between variables and ultimately, maximizing our insights of the data set and minimizing potential error that may occur later in the process.

Firstly, we imported all required python libraries we may need. This includes, **pandas**, **numpy**, **matplotlib** and **seaborn**. Data set was loaded into jupyter notebook using pandas. The **head** function was used to view some examples of the data set. We used the pandas **shape** function to know the number of rows and columns the data set has. Output was (4895, 15) meaning the data set has 4895 rows and 15 columns. We checked for the datatypes using the **dtypes** function. We noticed that most variable were saved as objects inexception of age that was save as float while signs of mental illness and body camara were Booleans. The data type of the age variable was changed to integer and the date variable was also changed to datetime datatype. The **column** function returned all the variables in the data set. Using the isnull function we got to know there were no missing values in our data set.

Knowing all the variables, we wanted to get a better understanding of the different values for each variable. To know the number of unique values each variable has, we used the pandas function, **nunique**. Output for ID was 4895 which equaled the number of rows while gender had 2 categories capturing ‘Female’ or ‘Male’. Likewise, manner\_of\_death variable had 2 categories captured as ‘shot’ or ‘shot and tasered’. The **describe function** summarizes the count, mean, standard deviation, min, and max for numeric variables. We found the minimum age to be 6 and the maximum to be 91.

Using the **unique function**, we looked at the discrete variables, including ‘arms\_category’, ‘threat\_level, ‘race’, and armed ’.The arms\_category variable has plenty category or values that was not distinctive. For example, a nail, a shovel and a toy weapon and a gun were classified as piercing object, blunt instrument, unusual object and guns respectively. We decided to classify the arms\_category variable by classifying unarmed and unknown as Unarmed. Piercing, sharp and hand tools were classified as sharp objects. The rest which include guns, vehicles, explosives, multiple etc. were put in the armed category. The threat\_level variable was also normalized to attack and undetermined values. The value ‘others’ in the original data set was classified as undetermined for the threat\_level variable.

There were no redundant columns or rows in the US shootings data set. The ID column in our view was irrelevant so we dropped the column and renamed the armed variable to type\_of\_arm for readability.

To perform our analysis as well as visualization of the data we did set some hypothetical statements to guide our analysis of the dataset. These include:

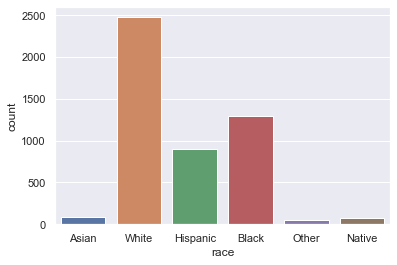
1. More blacks are shot than any race.
2. Youth are more likely to be shot than the older people.
3. People are more likely to be shot when they are armed.
4. Those who were shot showed signs of mental illness.
5. Victims are most likely shot because they attacked the policeman.
6. Victims that were fleeing were more likely to be shot.
7. Policemen with body camera were less likely to shoot their victims.

These statements were then turned into research questions and are presented below:

1. Are blacks more likely to be shot than any other race?
2. Are youths more likely to be shot than the older people?
3. What is the possibility that people are more likely to be shot when they are armed?
4. Do the police still shot victims who show signs of mental illness?
5. Were victims shot because they attacked the policeman?
6. Were victims shot because they were fleeing?
7. Are the policemen less likely to shoot when he wears body camera?

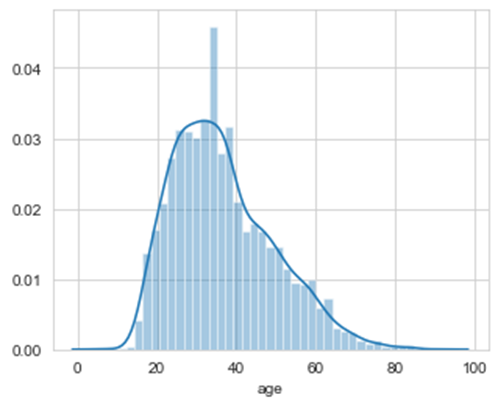
For the data visualization, we performed both univariate and bivariate analysis. These are presented below.

Surprisingly, we discovered that more whites were victims than any other race. They were followed by the black race. This was achieved using the value\_count function on the race variable. With this, our first hypothesis is solved. We used a count plot to generate the better pictorial view of the race variable as shown in (Figure 1)



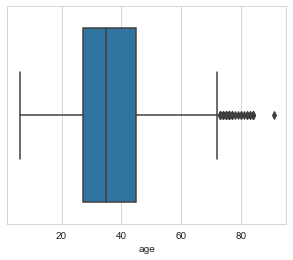
**Figure 1: Distribution of race of victims of shooting**

To graphical check for the distribution of age, the distplot command of the seaborn package was used. This distribution of age is shown in Figure 1. The distribution shows a positive skewness of the data implying that more victims were in their youthful age. As shown majority of the victims were below 40 years. This answers the hypothesis of the youth being the victims of police shooting in the US.



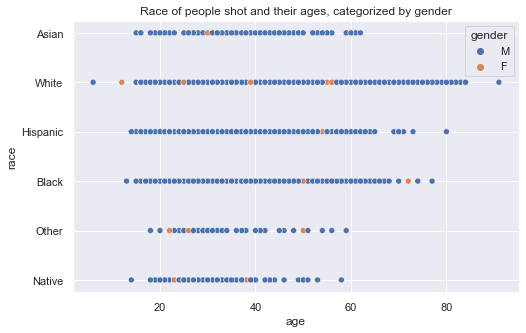
**Figure 2: Distribution of age of shooting victims**

It was obvious that there were outliers in the age variable. Hence a box plot was generated to confirm these outliers. This is presented in Figure 3. The figure shows clearly that there are outliers at older ages just above approximately 72 years.



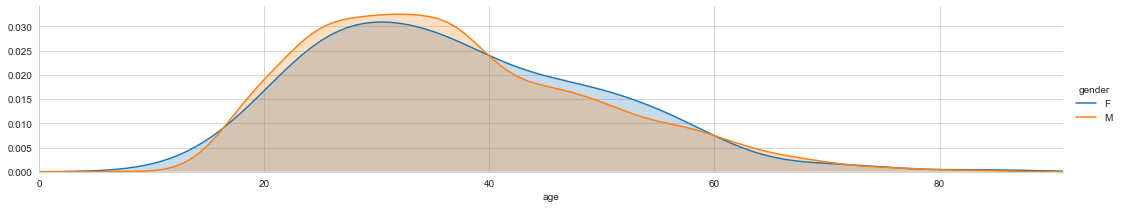
**Figure 3: Boxplot showing outliers of age of victims**

We went on trying to find the relationship between race and age grouped by gender using a scatter plot. Obviously from the chart (Figure 4) there is no correlation except for the following observations. More Males than females were shot, more whites were shot, and outliers were discovered. Both the minimum and maximum ages were of the white race.



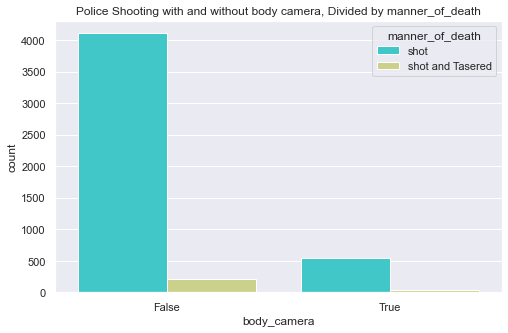
**Figure 4: Race of people shot and their ages, categorized by gender**

Using a kdensity plot we also discovered some interesting results of age and the sex of the victims. This presented in Figure 5. The figure shows that more proportion of women below 18 years were killed compared to their male counterparts of the same age. This was also true for ages between 40 and 60 years.



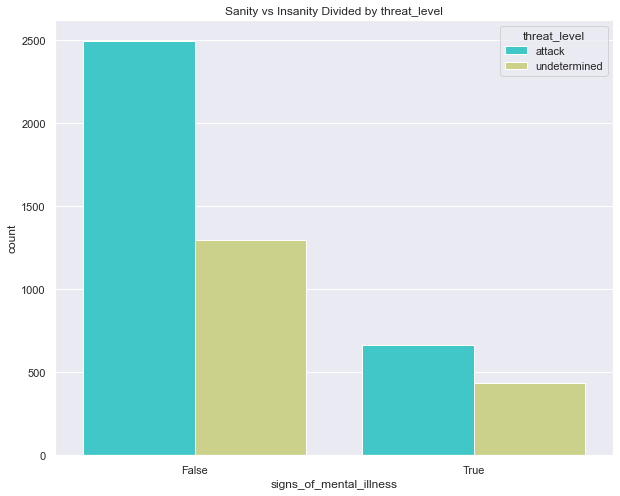
**Figure 5: Proportional distribution of age of victims by sex**

In trying to discover whether policemen with body camera were less likely to shoot their victim, we used a count plot to generate Figure 6. Truly, when the police are being watched they were more careful. Thus, according to Figure 6, most of the shooting were carried out by policemen who had no body camera on them compared with their counterparts who did have body camera on them.



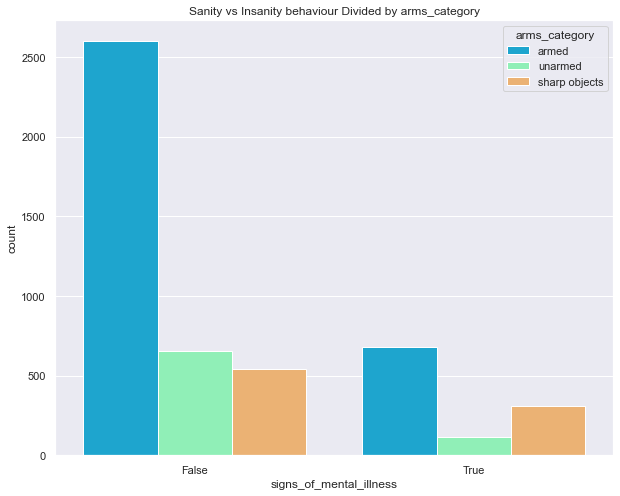
**Figure 6: Police Shooting with and without body camera, by manner of death**

To probe further, we wanted to know the actions of victims with no sign of mental illness that got them killed and compared to the actions of victims that showed no sign of mental illness. From Figures 7 to 9 persons showing signs of mental illness and those who showed no sign of mental illness exhibited very similar actions that led them to being shot. Figure 7 shows that most victims whether showing signs of sanity or not tried attacking the policeman who shot and killed them.

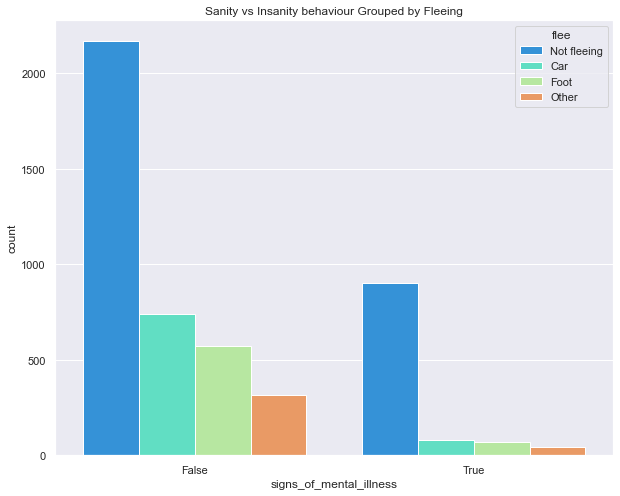


**Figure 7: Distribution of sanity of victims by their threat level**

Figure 8 also shows that majority of both sane and mentally unstable victims who were shot were armed. Interestingly most of the shot victims were not fleeing as also shown in Figure 9. However, for those who showed no sign of mental illness and tried fleeing, majority attempted fleeing by car followed by foot. From these comparisons, one can conclude that in shooting cases of these sort, in the heat of the moment, mentally sound individuals can become mentally unstable and may exhibit signs of a mentally ill person. We can also conclude that an armed individual is less likely to attempt fleeing the scene than an armed one.



**Figure 8:** **Distribution of sanity of victims by whether they were armed or not**



**Figure 9:** **Distribution of sanity of victims by whether they were fleeing or not**

Most shootings occurred in the more populated states in the United States. California with the highest populations being the highest with 701 shootings, followed by Texas with 426 shootings.