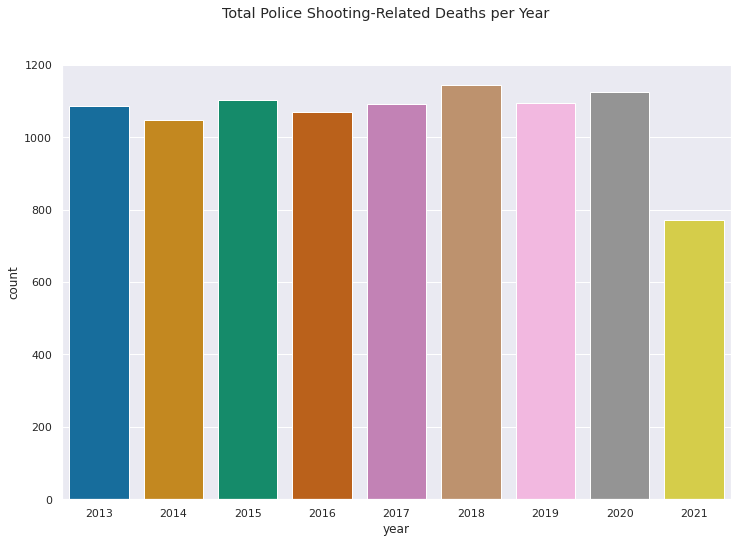
**Introduction**

The use of excessive force by law enforcement is a significant concern in the United States. This issue has been the topic of much discussion and debate in recent years, as it can lead to severe injury or death and has a profound impact on communities. To gain a better understanding of police brutality, it is crucial to analyze data on deaths caused by law enforcement. This project aims to examine a data set of deaths caused by police in the United States to uncover any patterns or trends that may give insight into the subject of police brutality. The data will be cleaned and modified to ensure its quality and consistency, then visualized and examined to extract meaningful information.

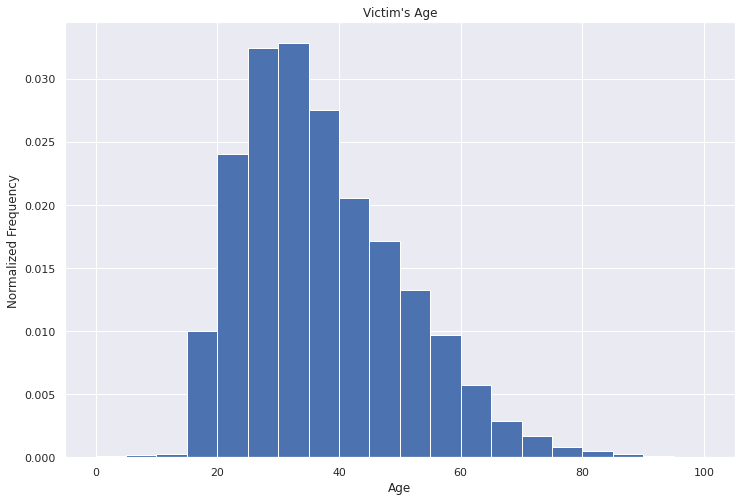
**Clean and transform the data**

In the data cleaning and transformation phase of the Exploratory Data Analysis (EDA), the following steps were taken to prepare the data for analysis:

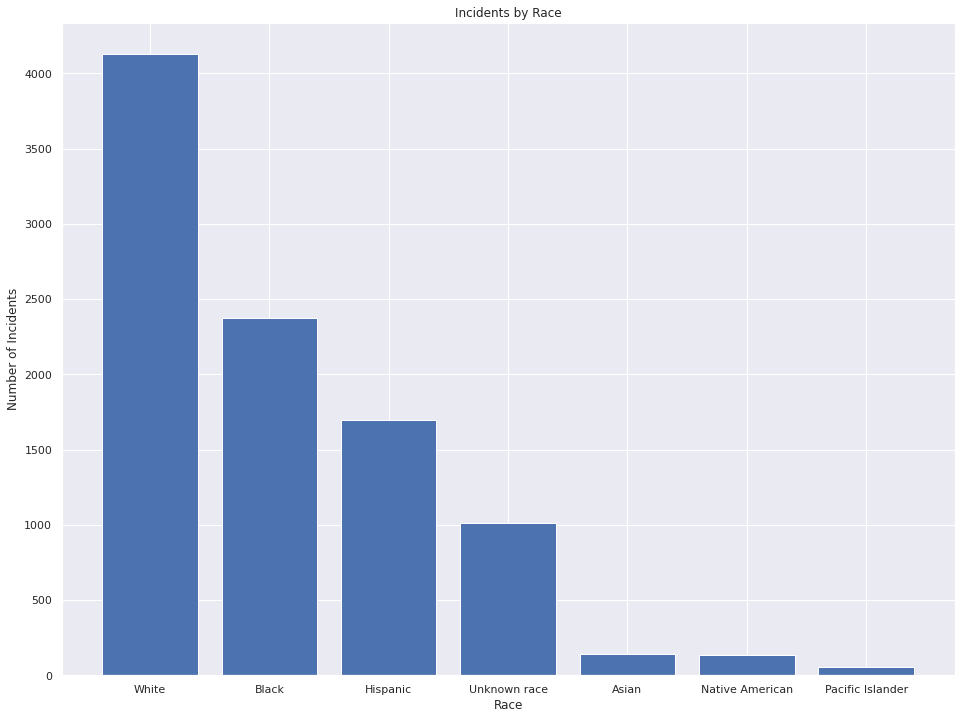
The "Victim's age" field was transformed so that it could be used effectively. This included converting it to a numeric format, replacing any invalid entries with a placeholder value, and then converting the field to an integer data type. The "Victim's gender" column was standardized by replacing any variations of "Male" with a consistent value. The "Armed/Unarmed Status" column was also standardized by replacing any variations of "Allegedly armed" with a consistent value. The "Symptoms of mental illness?" column was standardized by replacing any variations of "Drug or alcohol use" and "unknown" with a consistent value. The "Date of Incident (month/day/year)" column was converted to a specific data type and new columns were created to separate the year, month, and day of the incident.

**Data Exploration and Visualization**

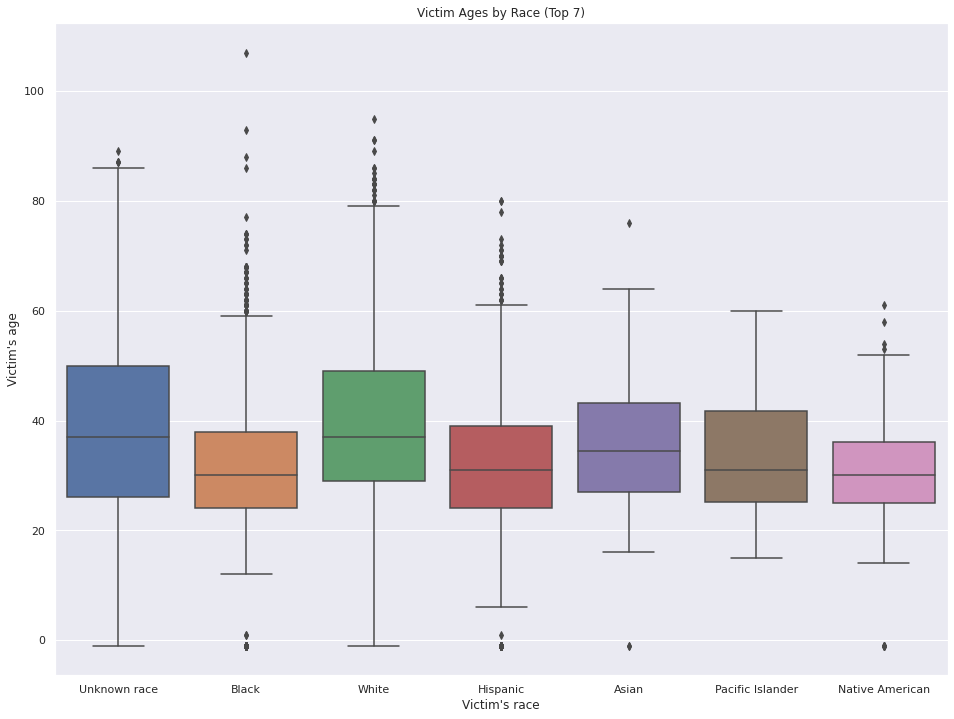
I observe that the visualization of deaths by year shows that the number of deaths in the data set has been relatively consistent over time, with small fluctuations from year to year. I note that the fact that the number of deaths has not been significantly impacted by the COVID-19 pandemic may suggest that the restrictions implemented in the United States to mitigate the spread of the virus have not had a significant effect on the number of deaths in the data set. Also, reason for the decline in 2021 is that 2021 does not end while data is being collected.



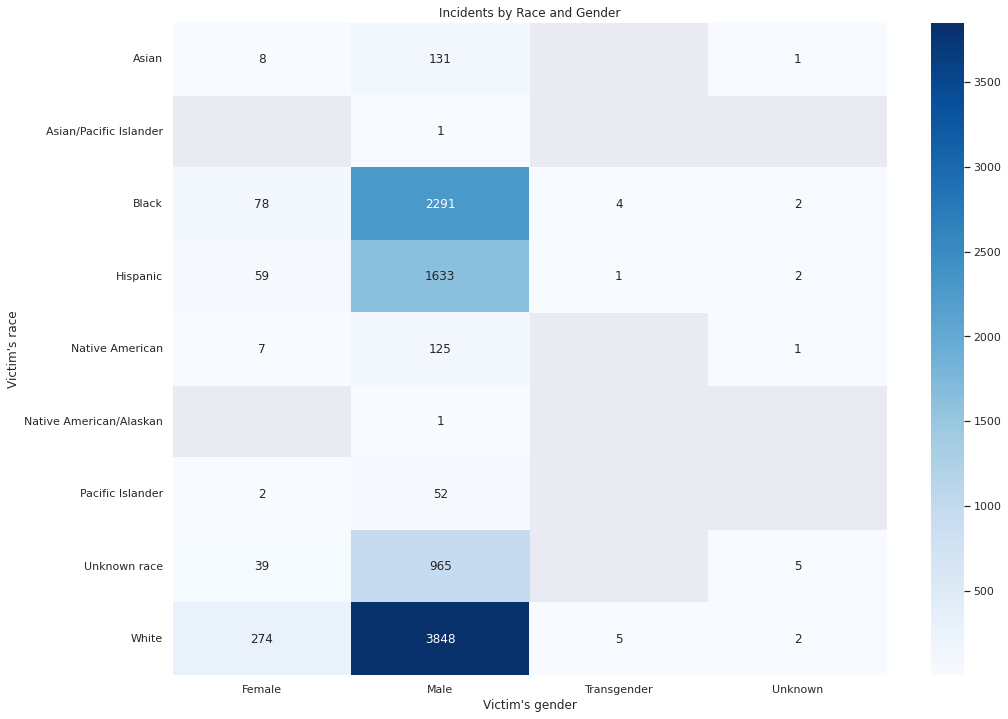
I observe that the majority of deaths in the data set, as represented in the age histogram, occurred among individuals between the ages of 20 and 40. I infer that this may be due to the fact that individuals within this age range may be more likely to be involved in situations that can lead to death, such as accidents or violent incidents.

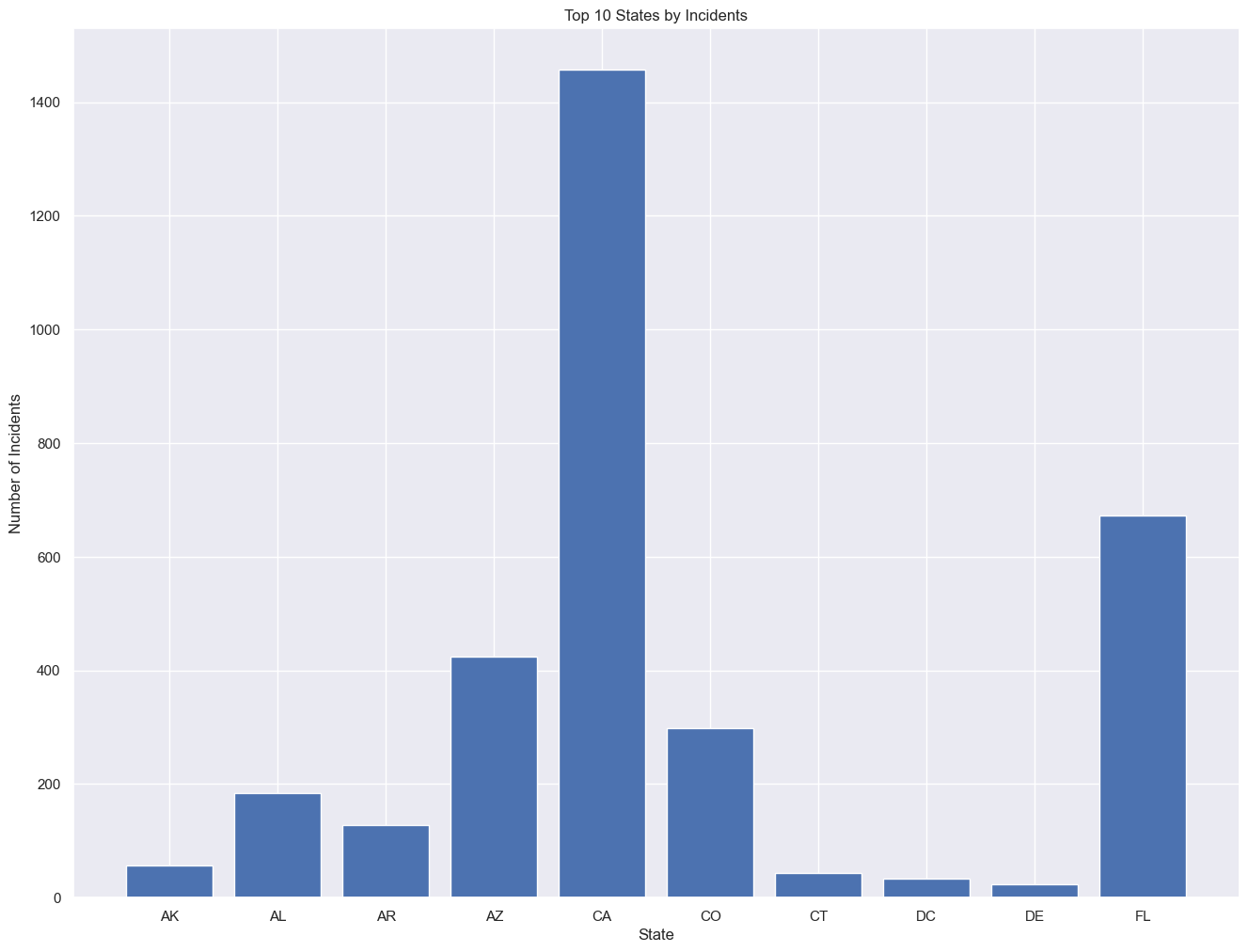


I observe that based on the visualization of the count of each race in the data set, the majority of the individuals in the data set are white, followed by black, Hispanic, and unknown. I note that there are fewer individuals in the data set who are Asian, Native American, and other races.

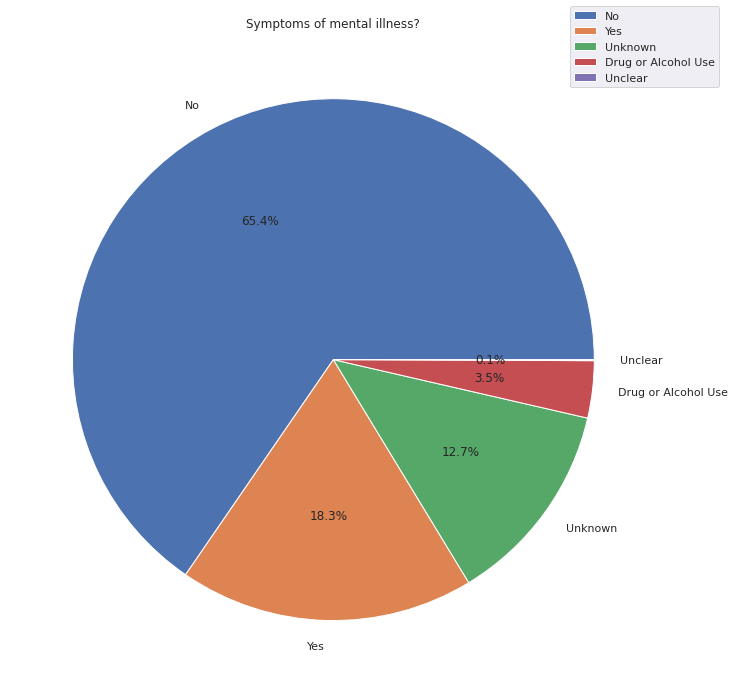


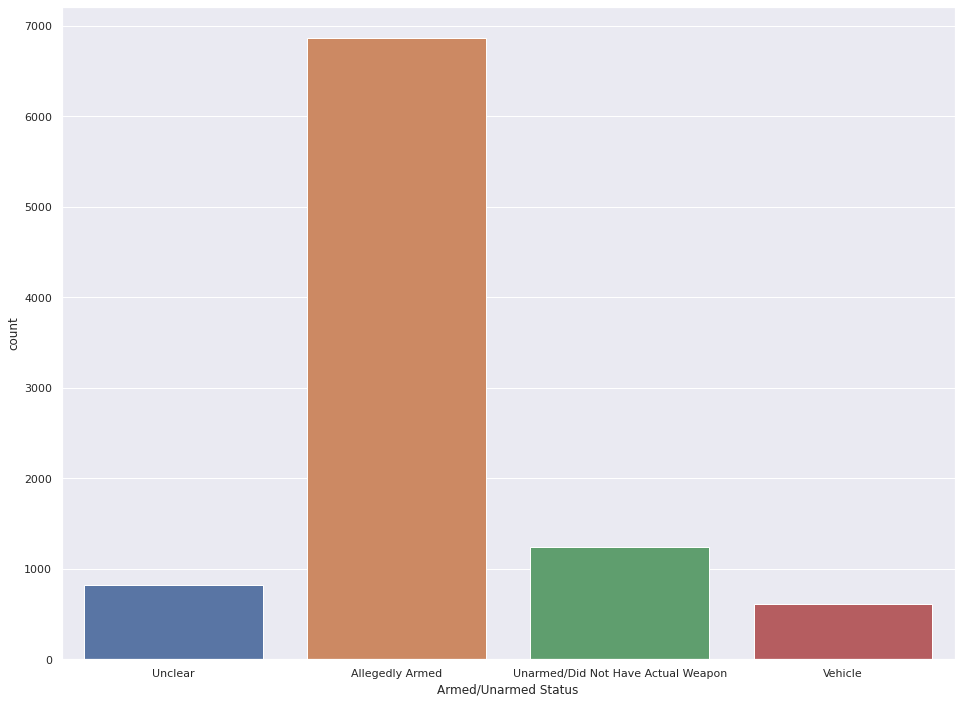
I observe that based on the box plot visualization of gender and age, the distribution of ages appears to be similar across all racial groups. This may suggest that there is not a significant difference in the age distribution of individuals in the data set by race.

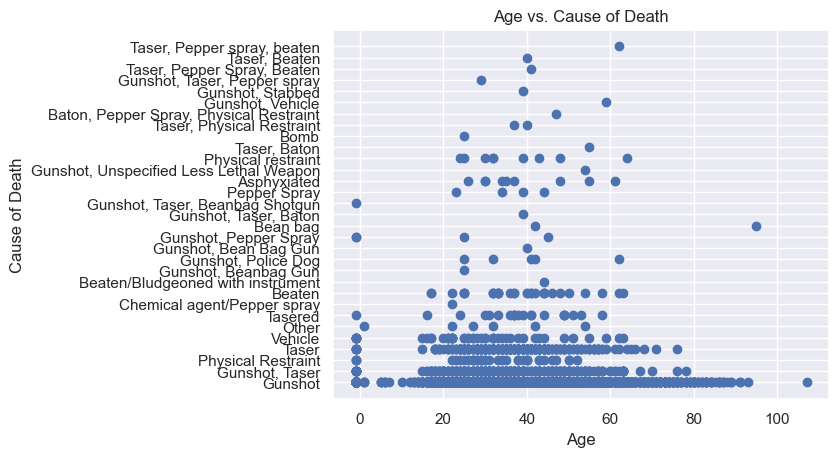


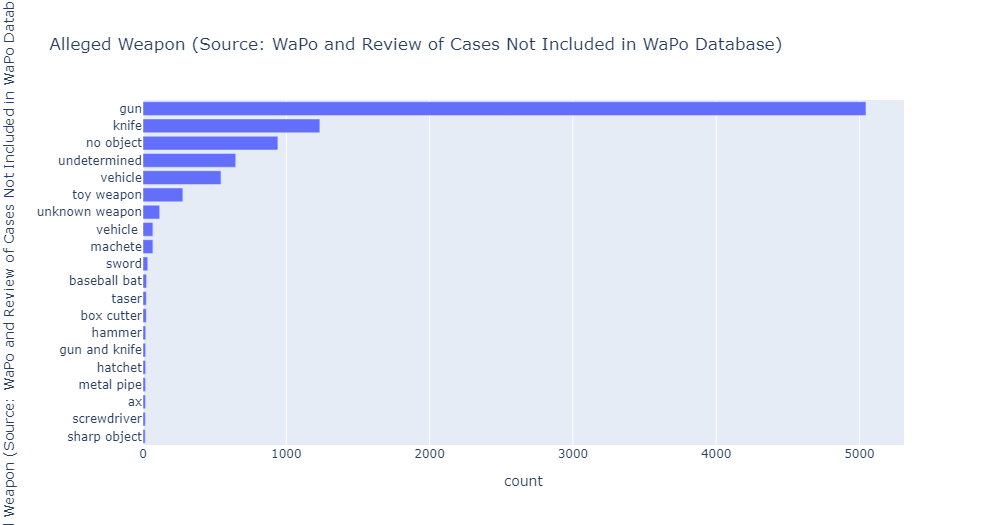
I observe that based on the visualization of gender in the data set, the majority of individuals in the data set appear to be male. This is consistent across all racial groups, with males making up the majority in each group. This finding may be due to a variety of factors such as the demographics of the areas where the incidents took place, societal norms and expectations regarding gender roles, and the prevalence of certain industries or activities that are more commonly associated with males.

I observe that based on the bar chart, the majority of deaths in the data set appear to have occurred in California, followed by Florida, Arizona, and Colorado. This may be due to a variety of factors such as the population size and demographics of these states, the prevalence of certain industries or activities that may be associated with an increased risk of death, or the presence of other risk factors such as a high rate of gun ownership.

I observe that based on the pie chart of the Symptoms of Mental Illness column in the data set, the majority of individuals (65%) did not have symptoms of mental illness at the time of the incident. This may suggest that the presence of mental illness is not a significant factor in the majority of incidents in the data set.



I observe that based on the visualization of the Armed/Unarmed Status of individuals in the data set, the majority of deaths appear to have occurred when the individual was allegedly armed at the time of the shooting. This may suggest that the use of deadly force by police is more likely to occur when the individual is perceived as posing a threat due to being armed.

I observe that based on the visualization of the Cause of Death column in the data set, the majority of deaths in the data set appear to have been caused by gunshot, taser, vehicle, or physical restraint. This may suggest that these are the most common causes of death in incidents involving police.

I observe that based on the visualization of the Alleged Weapon column in the data set, it appears that the most common alleged weapon in the data set is a gun, followed by a knife. There are also a significant number of cases in which “no weapon” was used, the weapon was undetermined, or the weapon was unknown. I note that it is not surprising that guns are the most common alleged weapon in the data set, given the high prevalence of gun ownership in the United States. The United States has a higher rate of gun ownership than any other country, and guns are involved in a significant number of deaths and injuries each year

**Conclusion**

In summary, the data set appears to show that deaths caused by police in the United States have been relatively consistent over time, with small fluctuations.

The majority of deaths occurred among individuals between the ages of 20 and 40, and the majority of individuals in the data set are white, followed by black, Hispanic, and unknown. As of the year 2020, the majority of the population in the United States identifies as white, with this group making up around 58% of the overall population. The largest ethnic minority group is made up of Hispanic and Latino Americans, accounting for nearly 19% of the population. The second largest minority group is Black or African American, making up 12% of the population[[1]](#footnote-1). With this information, it is normal that whites are the majority in the data set, but the relative proportion of Hispanic and Black individuals in the data set compared to their proportion in the overall population is interesting. It could be that the proportion of these groups in the data set would be different if there were no unknown race data in the dataset.

The distribution of ages is similar across all racial groups, and the majority of individuals in the data set are male. The majority of deaths occurred in California, followed by Florida, Arizona, and Colorado. It is not surprising that California and Florida are among the states with the highest number of deaths in the data set, as they are also among the most populous states in the United States. The large population size of these states may contribute to their high numbers of deaths in the data set.

The majority of individuals (65%) did not have symptoms of mental illness at the time of the incident, and the majority of deaths occurred when the individual was allegedly armed.

In conclusion, the EDA of the data set has provided a preliminary understanding of the characteristics and circumstances of the deaths in the data set.

Dataset link: <https://www.kaggle.com/datasets/jamesvandenberg/us-police-shootings-20132020>

1. [*"Racial and Ethnic Diversity in the United States: 2010 Census and 2020 Census"*](https://www.census.gov/library/visualizations/interactive/racial-and-ethnic-diversity-in-the-united-states-2010-and-2020-census.html) [↑](#footnote-ref-1)