**POLICE SHOOTINGS IN THE UNITED STATES**

**—- Team S4 —-**

**Sai Harsha Kotha**

 Dept of Computer Engineering

San Jose State University

[saiharsha.kotha@sjsu.edu](mailto:saiharshakotha@sjsu.edu)

ID: 016254927

**Surya Teja Palle**

Dept of Computer Engineering

San Jose State University

[suryateja.palle@sjsu.edu](mailto:suryateja.palle@sjsu.edu)

ID: 016587363

**Saketh reddy**

 Dept of Computer Engineering

San Jose State University

[sakethreddy.awala@sjsu.edu](mailto:saiharshakotha@sjsu.edu)

ID: 016724838

**Shreya Kommera**

Dept of Computer Engineering

San Jose State University

[shreya.kommera@sjsu.edu](mailto:suryateja.palle@sjsu.edu)

ID: 016703921

**Abstract:**

According to what our team has read from various articles over the internet, The Federal Bureau of Investigation does not compile this data, despite Congress' 1994 directive to the Attorney General to develop and publish annual statistics on police use of excessive force. So, there were many pages on the internet that came forward and started collecting the data of various police killings that occurred. Those are: “The Guardian '', “The Washington Post”, “The Lancet”. Out of all those pages, “The Washington Post'' is the one that’s constantly working on the article and updating the data almost every day whenever an incident occurs and is open source. So, we decided to take the data that’s present in “The Washington post” article into consideration for our project. Our motive is to analyze the data that’s given by the article and to obtain information such as How have the number of shootings have changed over time. Find the estimated count of incidents.

**Outline:**

* **Introduction**
* **Literature Survey**
* **Implementation**
* **Dataset Visualization**

**Introduction**

In the United States, police officers fatally shoot about three people per day on average, a number that’s close to the yearly totals for other wealthy nations. But data on these deadly encounters have been hard to come by.

Police in the United States shoot and kill about 1,000 people every year, according to an ongoing analysis by The Washington Post.

This is largely because local police departments are not required to report these incidents to the federal government. Also compounding the problem: an updated FBI system for reporting data and confusion among local law enforcement about reporting responsibilities.

As part of its investigation, The Post began in 2015 to log [every person shot and killed](https://www.washingtonpost.com/nation/2019/08/09/what-weve-learned-about-police-shootings-years-after-ferguson/?itid=lk_inline_enhanced-template) by an on-duty police officer in the United States. Since then, reporters have recorded [thousands of deaths](https://www.washingtonpost.com/investigations/interactive/2021/police-shootings-since-2015/?itid=lk_inline_enhanced-template). In 2022, The Post updated its database to standardize and publish the names of the police agencies involved in each shooting to better measure accountability at the department level.

**Literature survey**

In a survey conducted by **Manhattan Institute colleague** **Eric Kaufmann**, for example, eight in 10 African-Americans and about half of white Biden voters said that they thought that young black men were more likely to be shot to death by police than to die in a car accident—one of the largest mortality risks to the young and healthy.

Another survey, by ***Skeptic* magazine**, showed that more than a third of liberal and very liberal respondents thought that the number of *unarmed* blacks killed by police each year was “about 1,000” or more. About a fifth of those calling themselves “very conservative” thought the same thing.

Yet another survey, from a **trio of academics**, found that about four in 10 African-Americans reported being “very afraid” of being killed by the police, which was roughly twice the share of black respondents who reported being “very afraid” of being murdered by criminals, as well as about four times the share of whites who reported being “very afraid” of being killed by the police.

The assumption of widespread, highly consequential police racism has also inspired hasty policy changes. For example, “implicit bias” training has become common for police officers, despite the fact that, as two policing researchers put it in 2018, “no empirical evidence exists on the impact of implicit bias training on officer decision-making in the field, whether officers who are trained in implicit bias are perceived to be fairer by citizens, which training modality (e.g., classroom vs. simulation-based) is most effective in producing persistent changes in police behavior, or how long training effects last.

**Implementation:**

We used a total of 4 datasets. Out of which 3 were from Kaggle and 1 from Washington post.

The first dataset has the below columns:

|  |  |
| --- | --- |
| **Column Names** | **Description** |
| Id | Unique ID of each incident |
| Name | Name of the person |
| Date | Date of incident |
| manner\_of\_death | How were they killed/died |
| Armed | Were they armed ? |
| Age | Age of victim |
| Gender | Gender |
| Race | What’s their race ? |
| City | City of incident. |
| State | State in which the incident took place |
| signs\_of\_mental\_illness | Where they mentally ill |
| threat\_level | Were they attacking ? |
| Flee | Were they fleeing ? |
| body\_camera | Did the police official have a body cam ? |
| longitude | Longitude of the incident. |
| latitude | Latitude of the incident. |
| is\_geocoding\_exact | Is it the exact location ? |

**:**

**The second dataset was from Kaggle: Police shootings 538.csv**

Second dataset has the following columns:

|  |  |
| --- | --- |
| **Column Names** | **Description** |
| Person | Name of the person |
| Dept | Dept location |
| Eow | Sorting it by End of the week |
| Cause | Cause of Death in detail |
| Cause\_short | Short detail how he is killed |
| Date | Date when the person got killed |
| Year | Year when the person got killed |
| Canine | Resemblance to dogs |
| Dept\_name | Dept Name |
| State | State of the Dept. |

Table 1: dataset-1 description

The next two datasets almost have columns and in resemblance with the above dataset.

By comparing the above four data sets we found that police shootings from 2015 to 2021 provide more reliable content to visualize the data because it has more columns and less erroneous data related to the type of incidents that occurred.As mentioned earlier, we will use linear regression, ARIMA model forecasting has been performed on the data set and we have used numpy, pandas, python, matplotlib, seaborn, and Jupyter notebooks.

**Data Preprocessing Steps:**

1. We calculated the null value percentage in each column and found that there are a maximum number of null values in the Race column.
2. In the name field we got to discover some of the missing values and instead of giving some random names, we just removed the name column as it won’t serve any model.
3. Even in the Race column we just went with the value called “Other”, because considering the race blindly doesn’t provide the accurate information.
4. Correcting the age column and keeping in between 0 to 100, as some of the values were in the negatives, so that the accuracy or prediction will be accurate according to the dataset.
5. Checking the date entry whether it is valid or not and putting it in a specified format (MM-DD-YYYY)

**DATASET VISUALIZATION:**

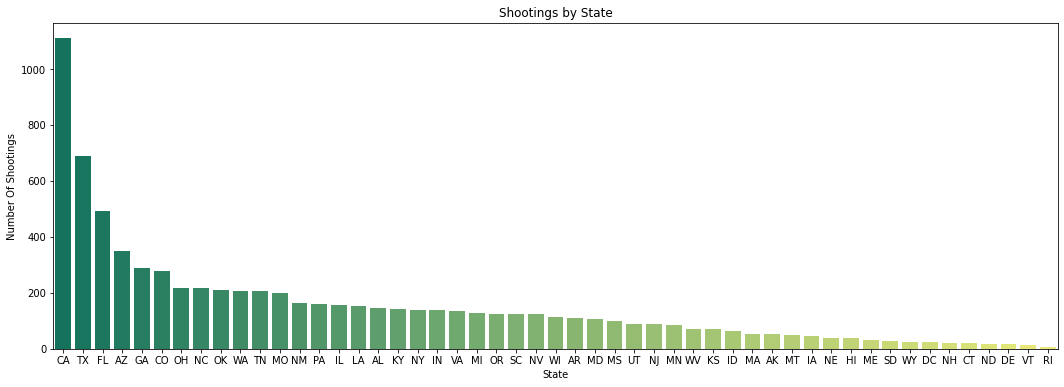
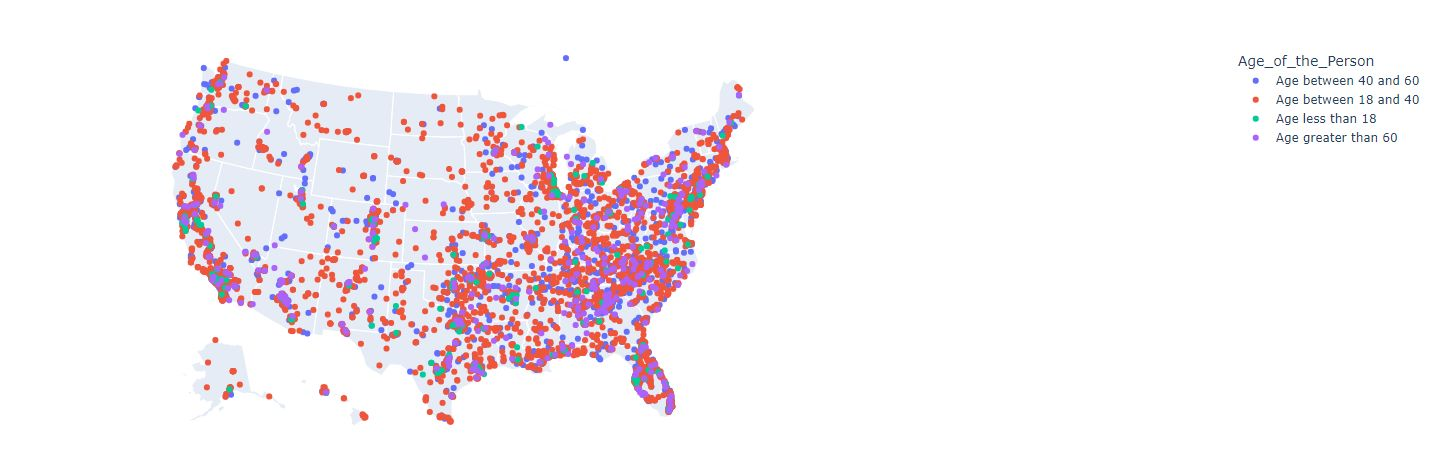


Fig: shootings by state

The above figure is the plot between the state and the number of shootings that happened in that state. The highest number of incidents happened in CA. The least happened in Rhode Island.

The below is the choropleth indicating the ages of the persons in different states involved in the police shootings in United States

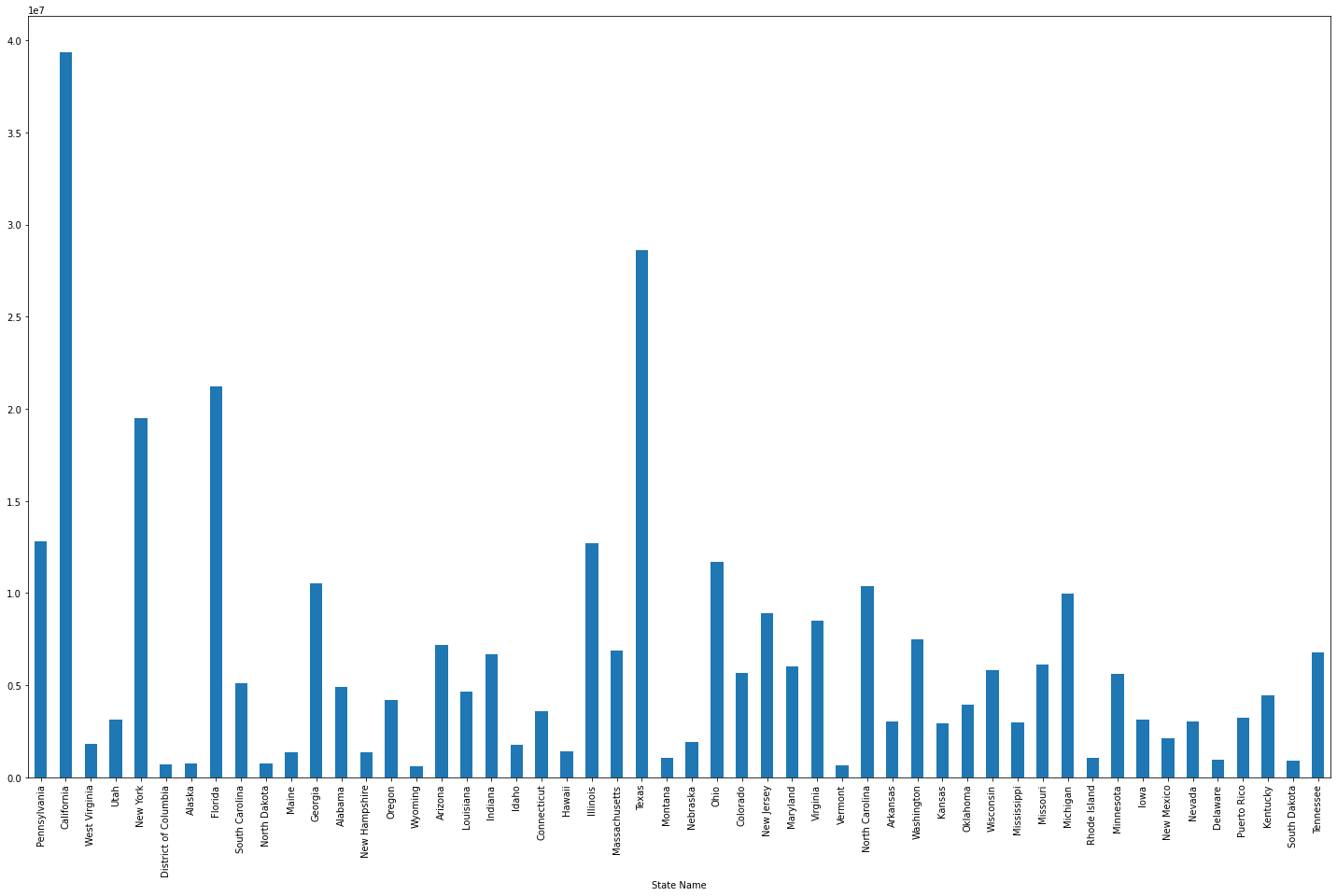




**CENSUS DATA:**

We requested for the API key from the census and got the key, thus we used the state wise and race wise data from the census and used it to perform various visualizations.

|  |
| --- |
| **url = 'https://api.census.gov/data/2020/acs/acs5?'**  **params = {'get' : 'NAME,B01001\_001E,B02001\_004E,B02001\_005E,B01001H\_001E,B01001B\_001E,B01001I\_001E',**  **'for' : 'state:\*'}**    **r = requests.get(url, params=params)** |

**** Fig: State name vs total population

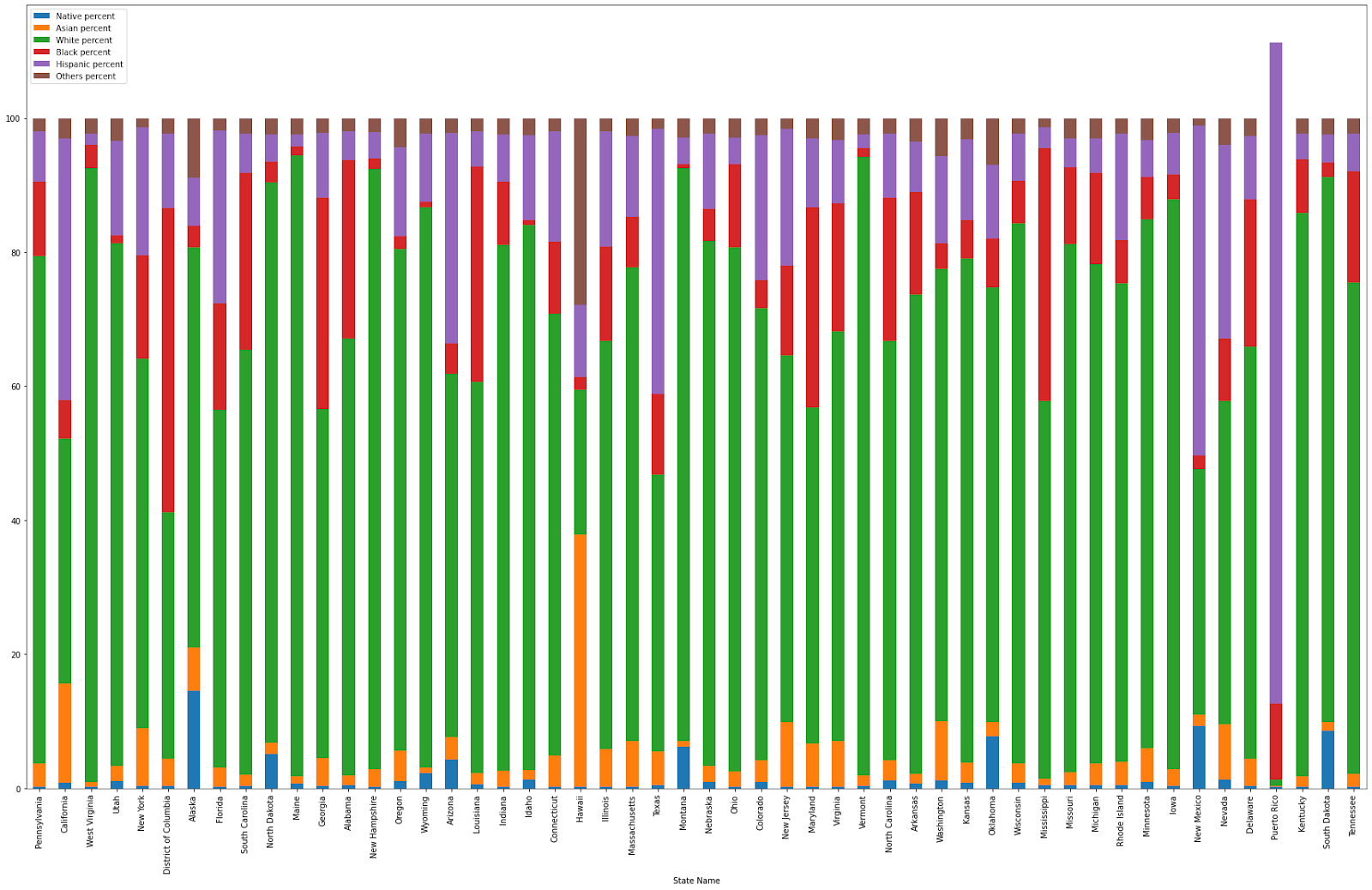
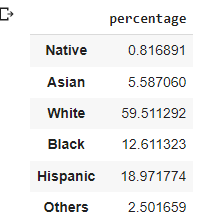


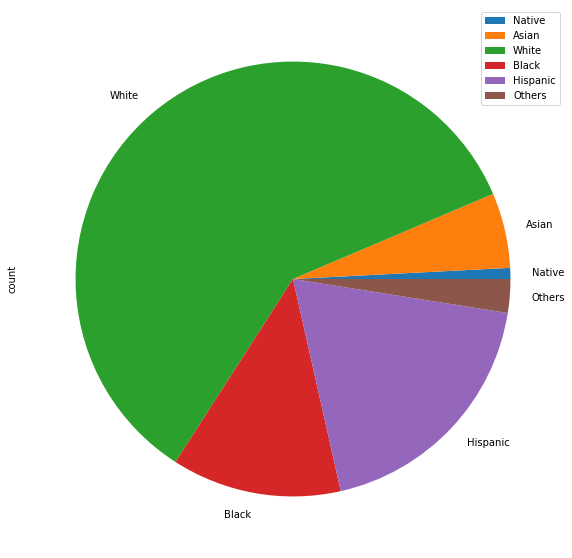
Fig: state vs percentage of people race wise

The above plot is plotted from the data we took from the census. It is clear that the most populated state is California, this explains us the reason behind numbers in the various plots.

From the above plot it’s clear that most of the states have white people more in number.

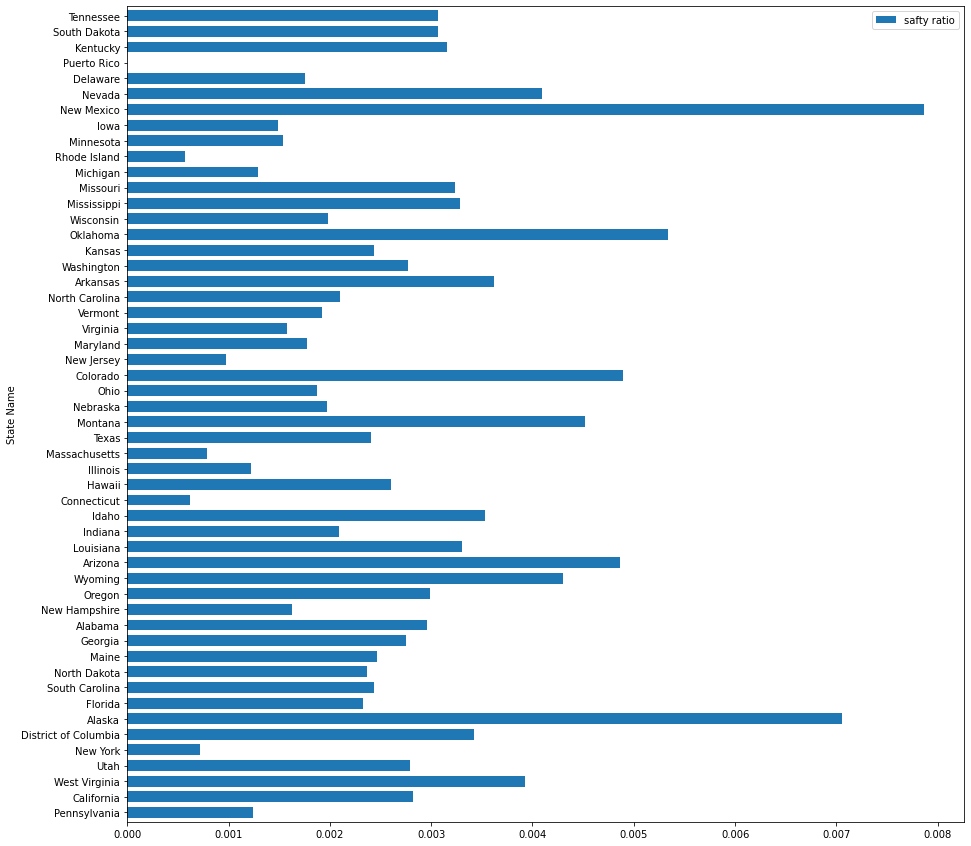


**Percentage of people by race**



Going further, comparing the number of incidents in each state to the population of the state to know the actual percentage of the population that get involved in the crimes.

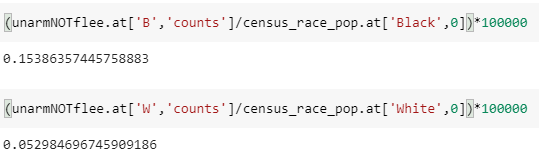
For example, the total population of **Pennsylvania is** 12794885. The number of incidents that happened in that state as per our dataset are 159. So, the rate becomes 0.001243.



From the above plot it's clear that, even though the number of incidents are more in California the chances for a citizen getting involved is less. New-Mexico is the most unsafe.

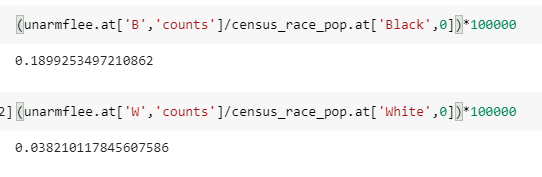
Took our original dataset and compared it with the census race wise population to know the chances of a person from black race getting killed compared to that of white in various scenarios.

**Person is not fleeing and unarmed:**

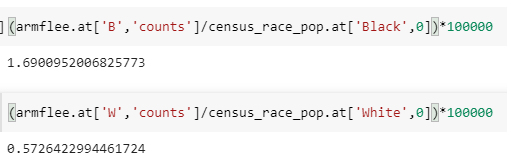
****

This is for the case where the person is not fleeing and unarmed, in this case the percentage of a person from **Black race getting killed is 3 times more than a white person getting killed**.

**Person fleeing and unarmed:**

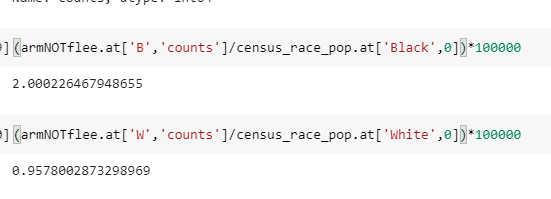
****

The chances of getting killed for a black man if fleeing is **6 times** more than for a white person.

**Person fleeing armed:**

The chances of getting killed for a black guy if fleeing is **more than 3 times** higher compared to a white person.

**Person Not Fleeing and is Armed:**



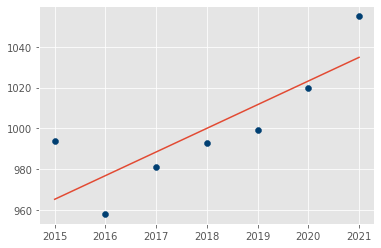
The chances of getting killed for a black guy if he is not fleeing and is armed are **more than 2 times** higher compared to a white person.

It’s very clear that, no matter what the circumstances are. **The chances for a person from black race have high chances of death** compared to any other race.

**Linear Regression:**

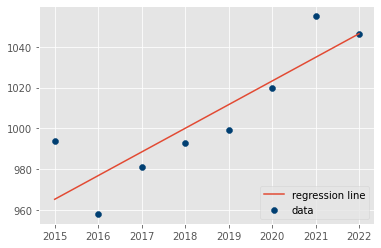
A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable. The differences between expected and actual output values are minimized by linear regression by fitting a straight line. We get a straight line/ best fit line by least square method.

In our project, from the original dataset “US Police shootings in from 2015-22.csv”. I extracted 2 columns namely year and count (number of killings). Year on the X-axis and count on Y-axis.



In the above image, the red coloured straight line is the ‘best fit’ line. With the help of this line we get the slope equation and thus find the dependent number.

‘*The variable you want to predict is called the dependent variable.*’



The predicted value for the year 2022 is 1047.

**Auto Regressive Integrated Moving Average (ARIMA):**

Auto Regressive Integrated Moving Average, is a set of models that explains a time series using its own previous values given by the lags (Auto Regressive) and lagged errors (Moving Average) while considering stationarity corrected by differencing (opposite of Integration.) In other words, ARIMA assumes that the time series is described by autocorrelations in the data rather than trends and seasonality.

**AR**: Autoregression.

**I**: Integrated.

**MA**: Moving Average.

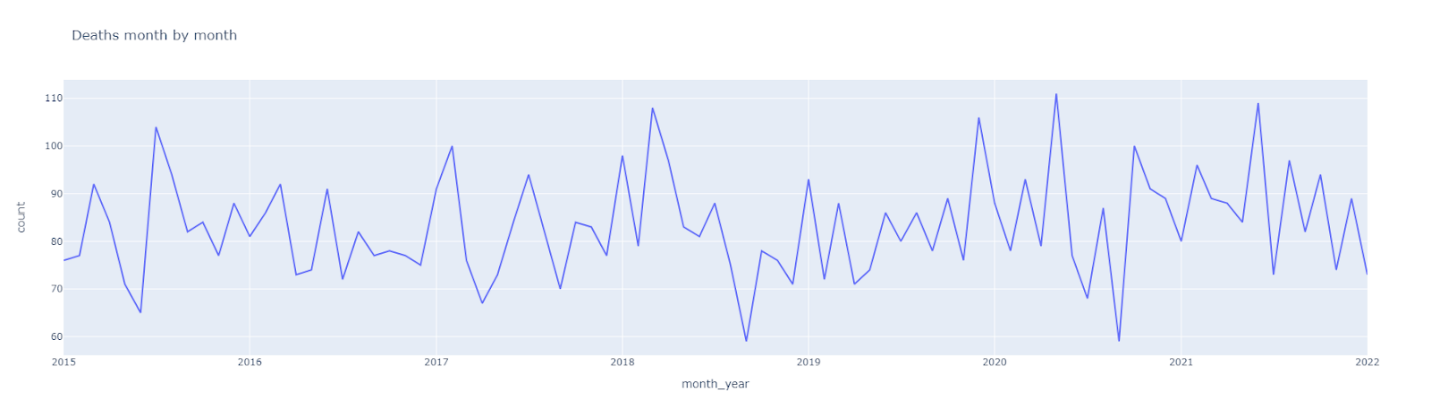
A standard notation is used of ARIMA(p,d,q).

p: Lag order.

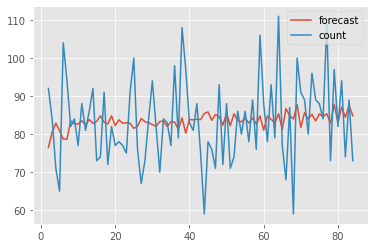
d: Also called the degree of differencing (The number of times that the raw observations are different).

q: Also called the order of moving average (The size of the moving average window).

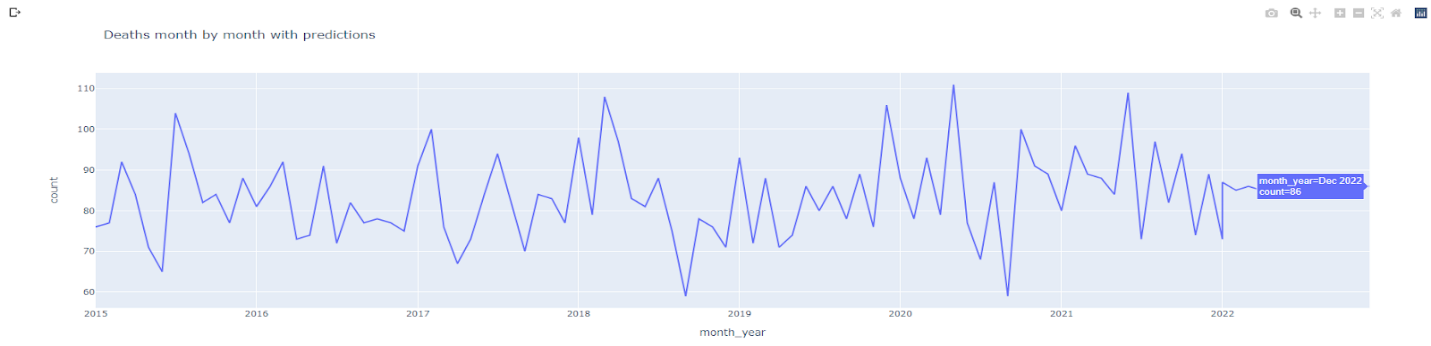
For our dataset we used (3,1,1) as this combination gave the best results.



The above figure is a plot between month\_year column and count (no.of killings).



The red colored waveform in the above plot is the prediction plot we got. When we forecast it for the next 12 months, we are going to predict the increased average mean of no.of killings per month.

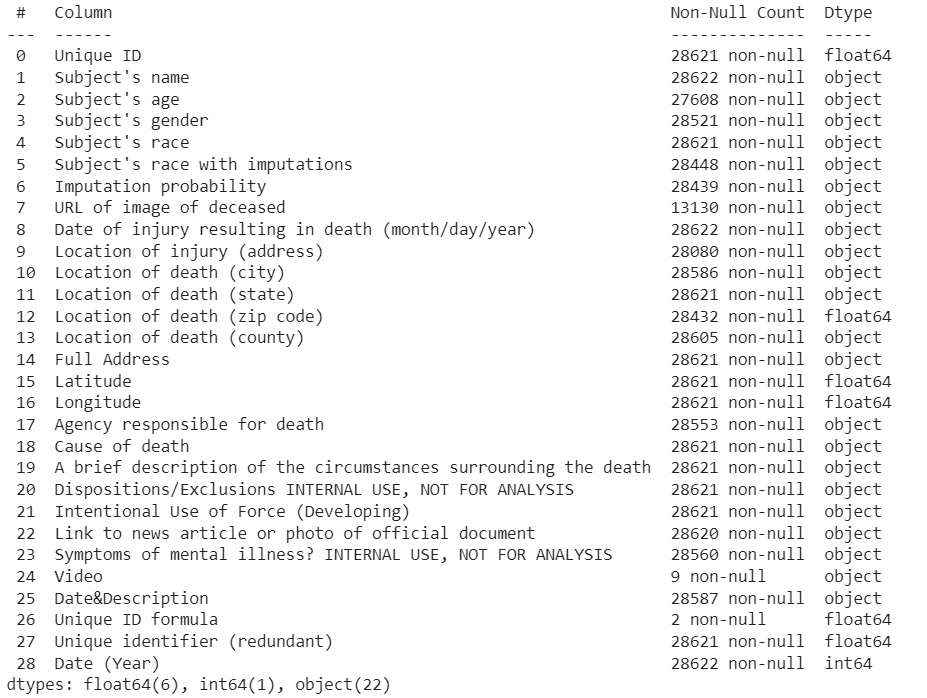


**Washington Post Police Shooting Analysis:**

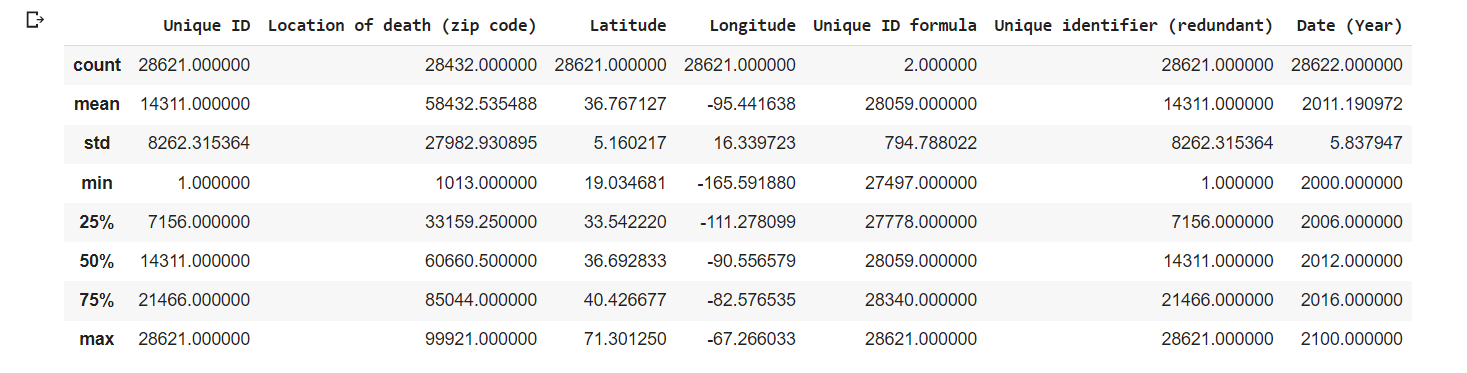
This research will examine murders committed by American police between January 2015 and February 2017 in order to better understand how such incidents are handled. The main objective is to just look at the data and determine if any trends are present.

**General Look at the Dataset:**

The dataset consists of the following columns. It contains 28 columns which has datatypes of float, int and objects.

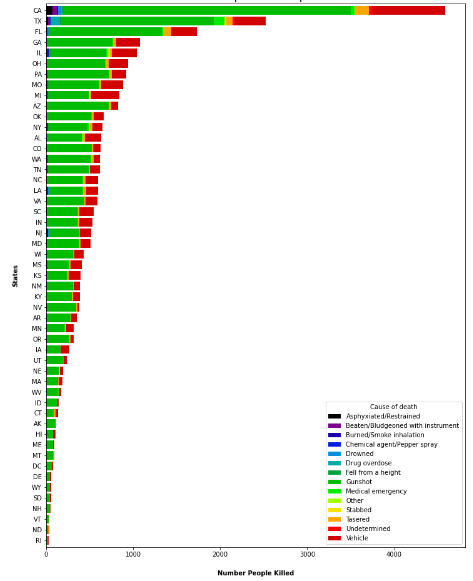


The describe method returns the following output for our dataset.



**Causes of People Death per State:**

Making a distinction by state, it is once more apparent that gunshot was the primary factor in the fatalities. After then, cars started killing people. The states with the most fatalities appear to be California, Texas, and Florida.



**Causes of People deaths per City:**

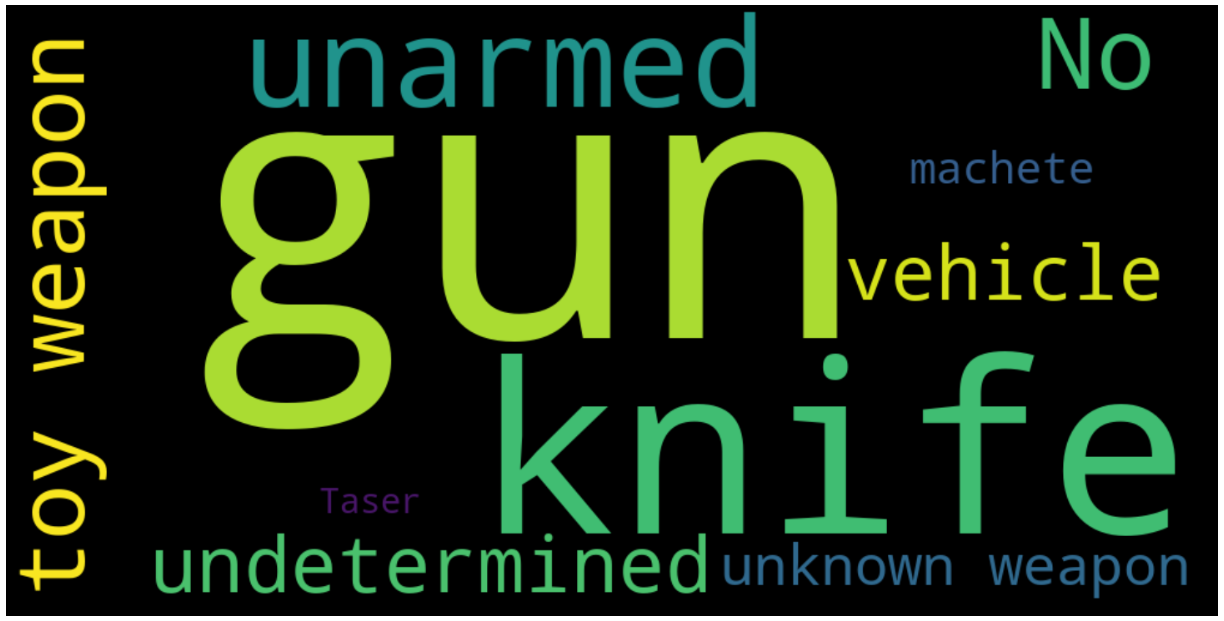
For the cities with the most murders, we created a word cloud. According to a review of the aforementioned graph for the top 20 cities, Chicago, Houston, and Los Angeles were found to be the bloodiest cities.

According to data up to 2016 and this data bounce up to 2020, Los Angeles and Houston were the bloodiest cities, followed by Phoenix and Las Vegas.



**Analyzing weapons used for killings:**

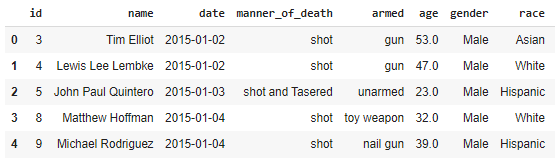
However, some also possessed knives or toy weapons. Guns were used by the majority of the victims. A lot of the time, no weapons were found, not even on the people.

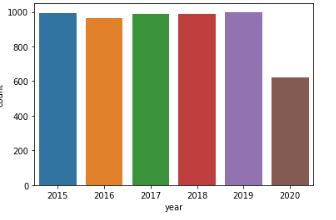


**Analysis of number of shootings using date factor:**

First we need to have preprocessed the date factor into one data format-YYYY-MM-DD. The processed dataset looks as shown below.

After that, we have used several plots to represent the data by year, month and day to know the number of killings as per it respectively.

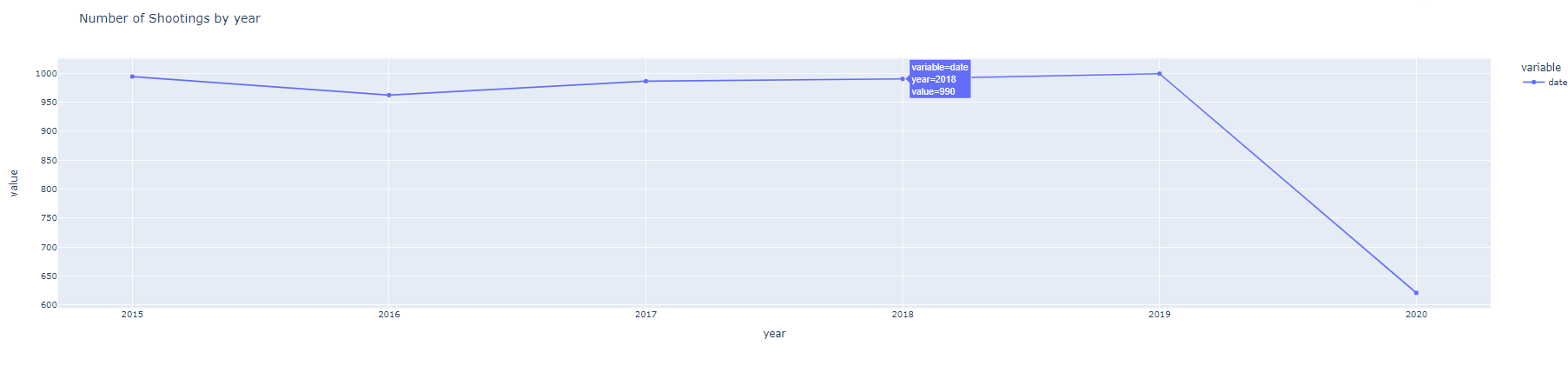




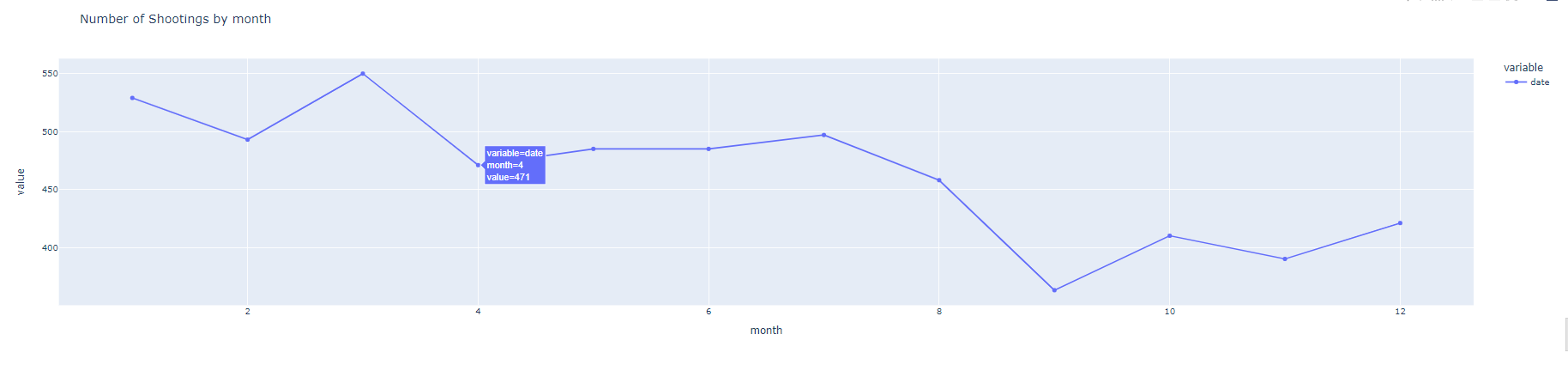
The above shown is the bar graph represented for the killings from the year 2015 to 2020 taking the count on y axis and year on x axis. Among them 2015 and 2019 years has the most number of shootings.

We have represented some interactive line graphs to understand much more on killings over years, months, days and as well as cumulative over all.

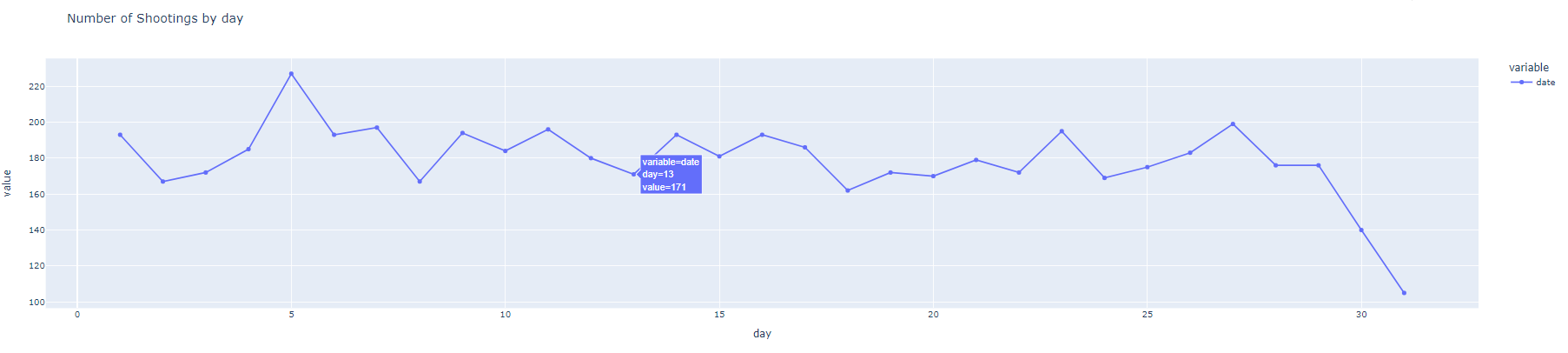
1. Number of shootings per year:



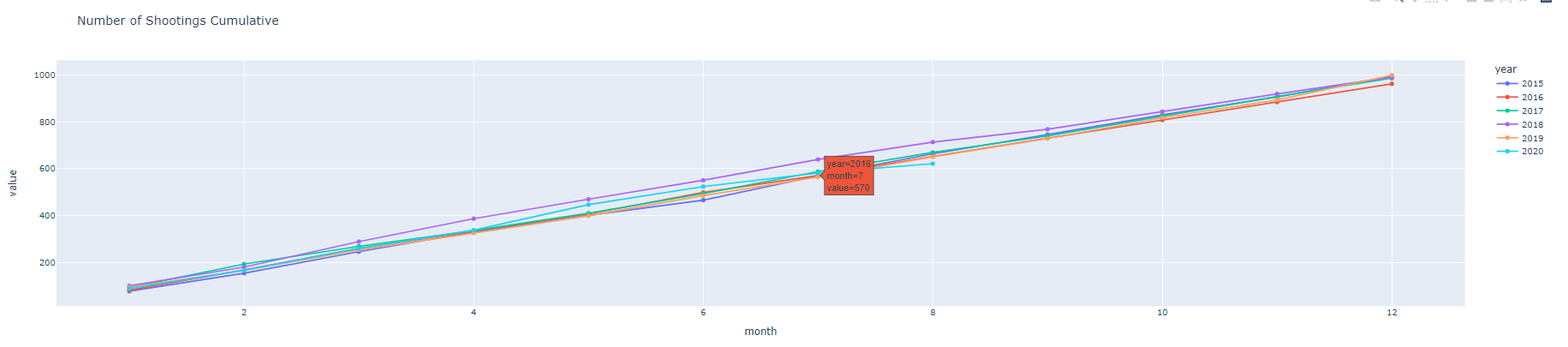
1. Number of shootings per month



1. Number of shootings per day



1. Number of shootings cumulative



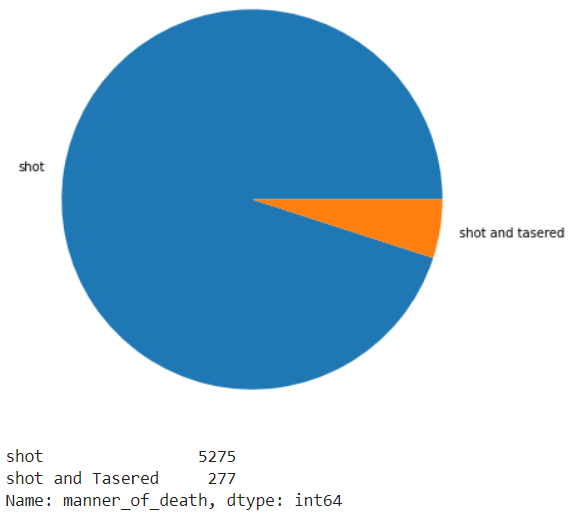
**Analysis of Pie Charts for Manner of Death:**

We have represented some pie charts which represents the manner of death using some set of labels.

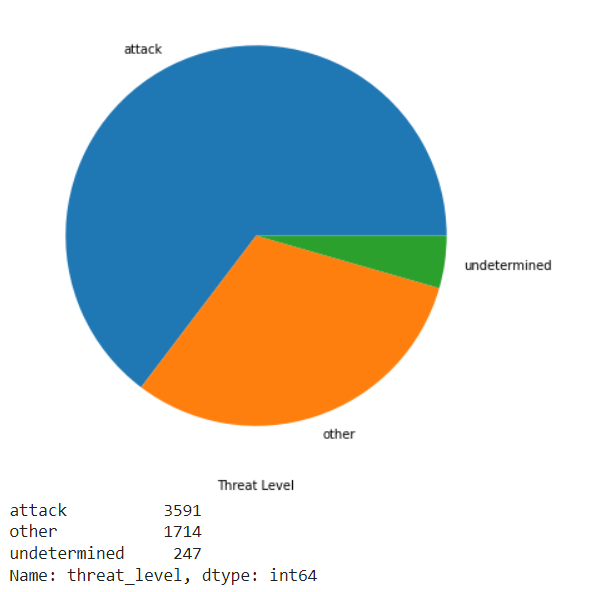
Using the labels ‘shot’, ‘shot and tasered’ we have represented a pie chart for number of shootings for these manner of deaths. From that we got to know that 5275 number of people have been shot whereas 277 number of people have been shot and tasered.

Using the labels ‘attack’, ‘other’, ‘undetermined’ we have represented a pie chart for number of shootings for these manner of deaths. From that we got to know that 3591 number of people have been attacked whereas 1714 number of people are of other category and tasered and 247 have been undetermined.

1. Using the labels shot and shot and tasered:

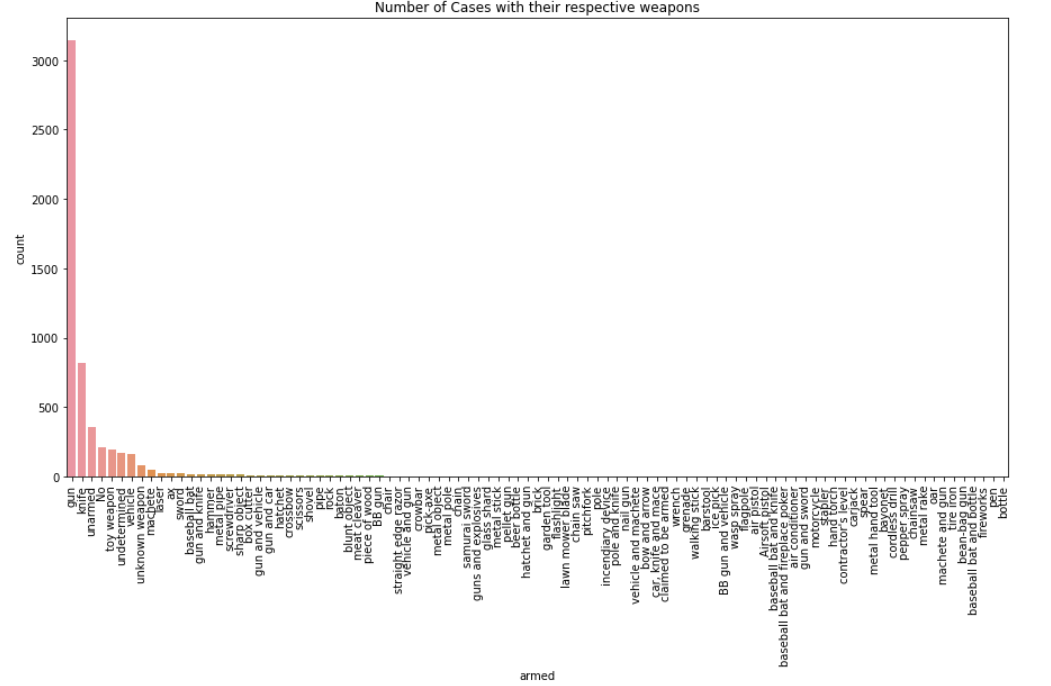


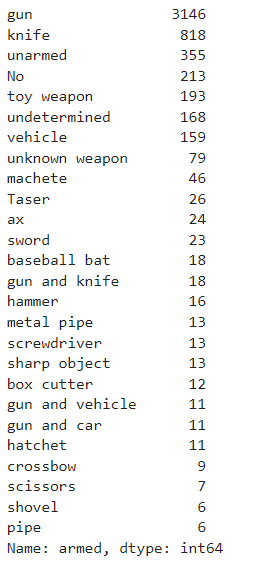
1. Using labels attack, undetermined and other:



**Number of cases with respective weapons:**

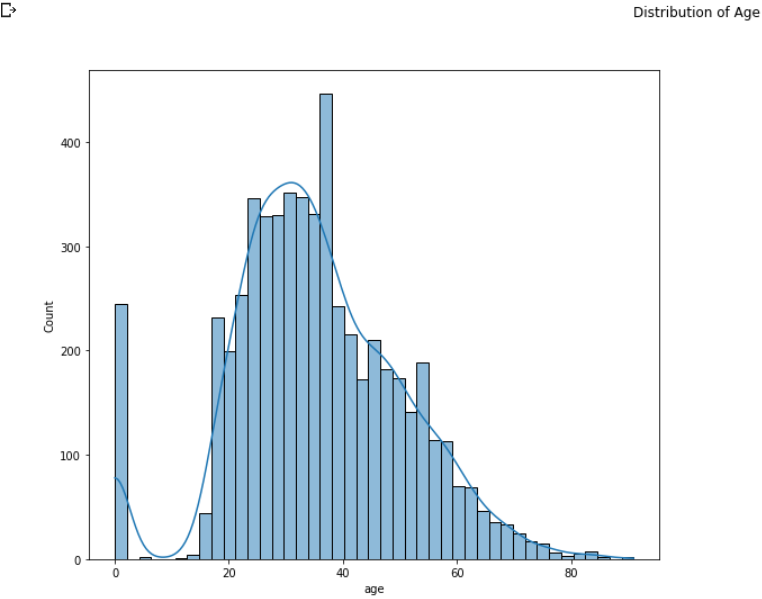
Below plot shows the cases registered with each weapon that the persons used during the police shootings in the United States. Most of the victims had guns, but some also had knives or toy weapons. Many times, there were no weapons present or none were even found on the people.

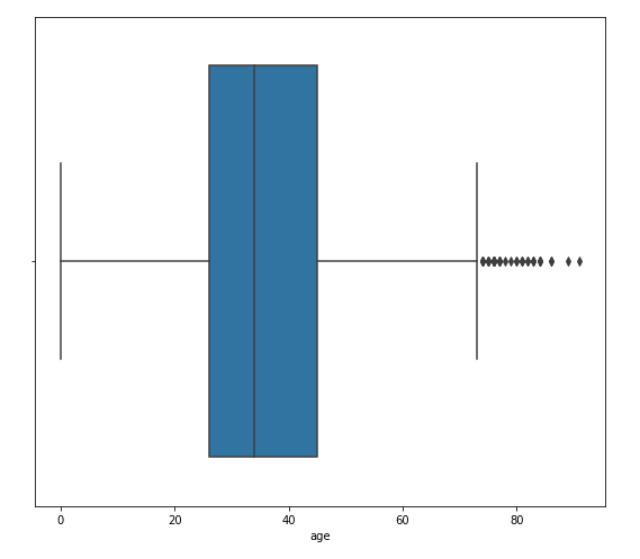




**Distribution of age:**

Let’s look at the distribution of age on x axis and count of shootings on y axis through bar plot and box plots. Most of the people's age lie between 20-45 as most of the people were young.

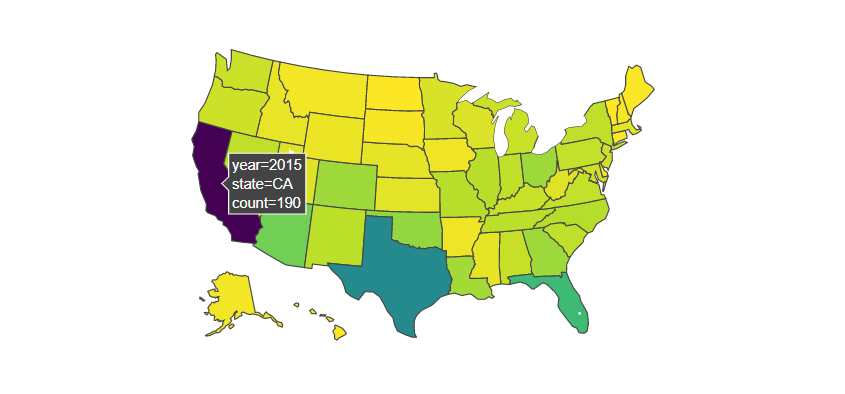


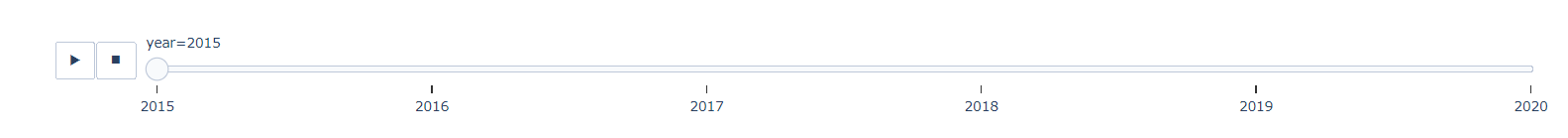


Most of the people's age who were involved in the police shootings lie between 20-45 as most of the people were young.

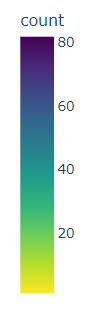
**Incidents Observed in Each State Over the Year:**

We have made a choropleth map in order to analyze the incidents that have been observed in each state in the United States. Choropleth maps display normalized data as shaded dots, lines, or areas using the Counts and quantities (Color) smart mapping symbol type. Choropleth maps provide information about your data.

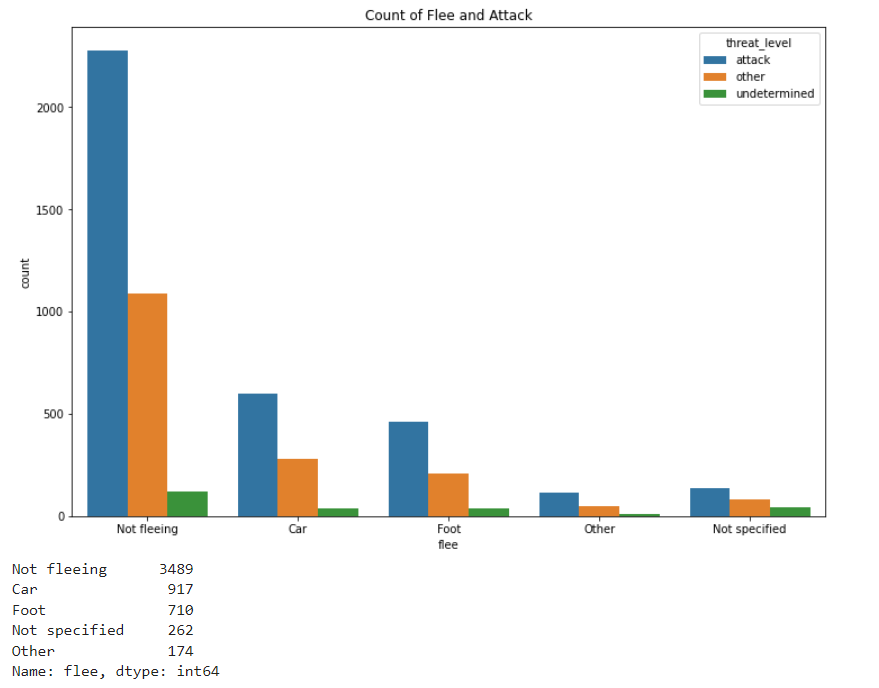




Among the years 2015 to 2020 the California State stands for the first position with the most number of incidents occurred.



**Count of Flee and Attack:**



From the above count-plot we observed that Most of the people tried to attack irrespective of their fleeing intention.