POLICE SHOOTINGS IN THE UNITED STATES

—- Team S4 —-

Sai Harsha Kotha
Dept of Computer Engineering
San Jose State University
saiharsha.kotha@sjsu.edu
ID: 016254927

Saketh reddy
Dept of Computer Engineering
San Jose State University
sakethreddy.awala@sjsu.edu
ID: 016724838

Surya Teja Palle
Dept of Computer Engineering
San Jose State University
suryateja.palle@sjsu.edu
ID: 016587363

Shreya Kommera
Dept of Computer Engineering
San Jose State University
shreya.kommera@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 by an on-duty police officer in the United States. Since then, reporters have recorded thousands of deaths. 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:

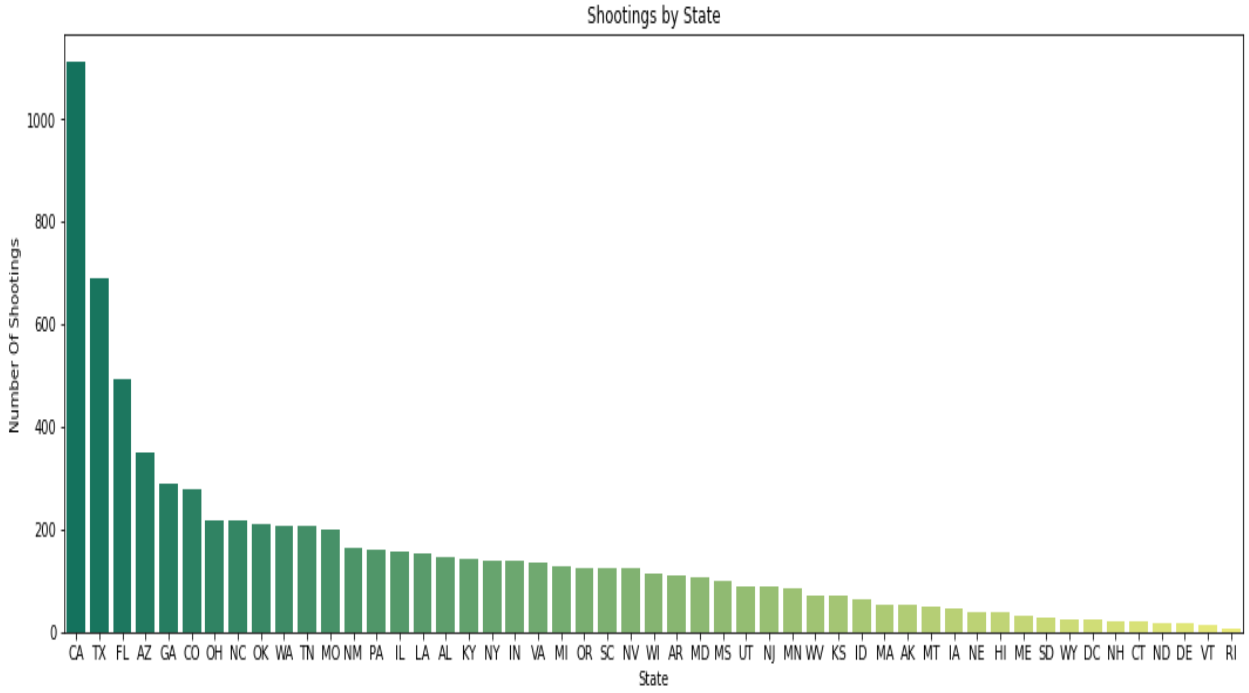


Fig: shootings by state

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)
```

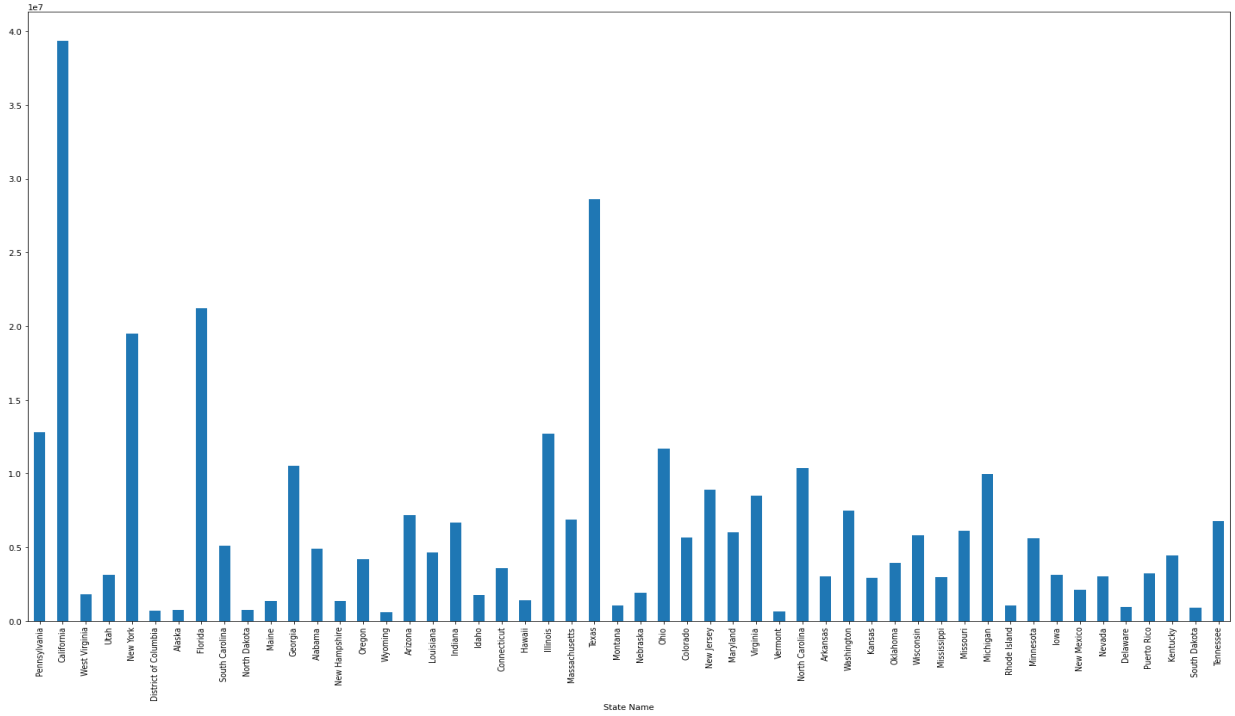


Fig: State name vs total population

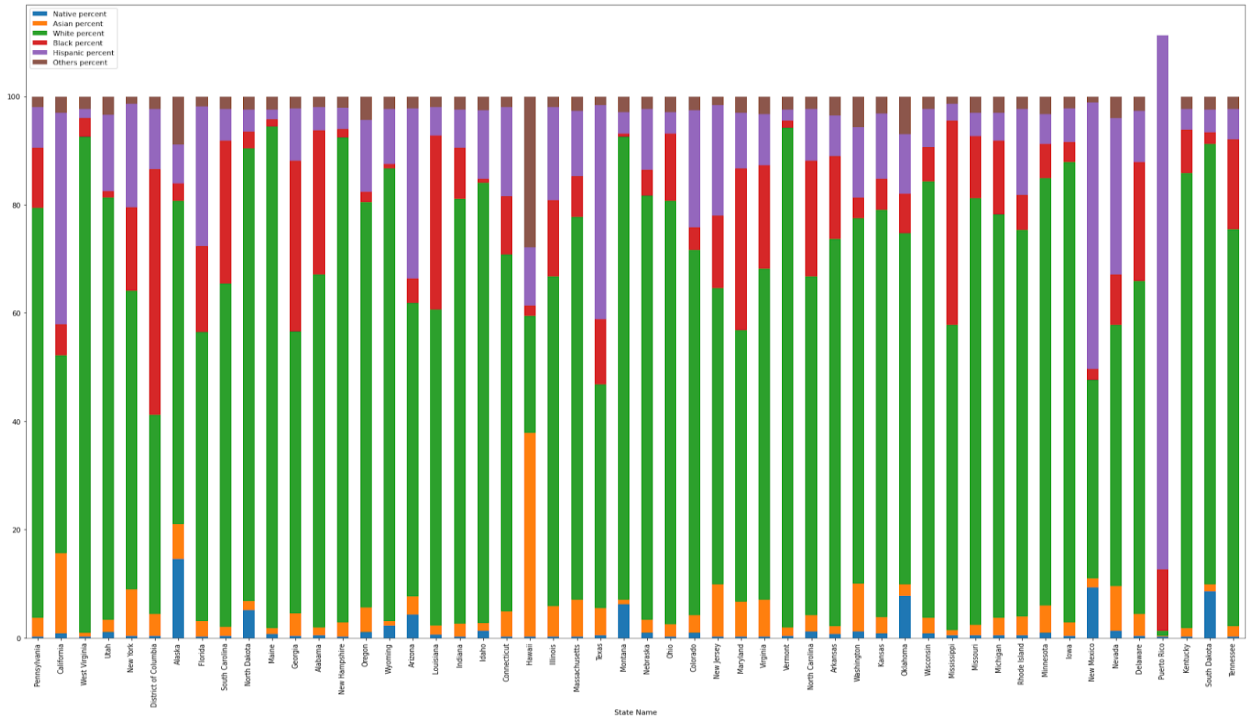


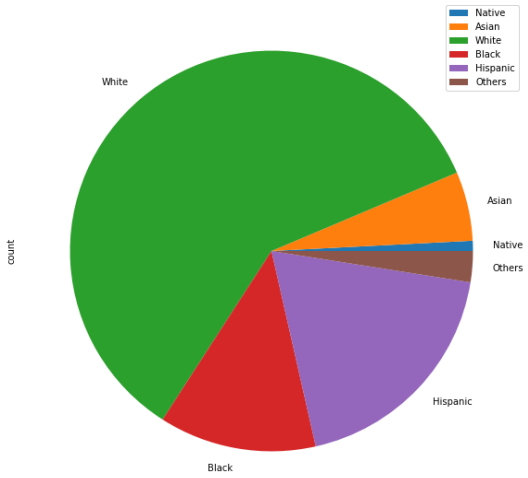
Fig: state vs percentage of people race wise

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.

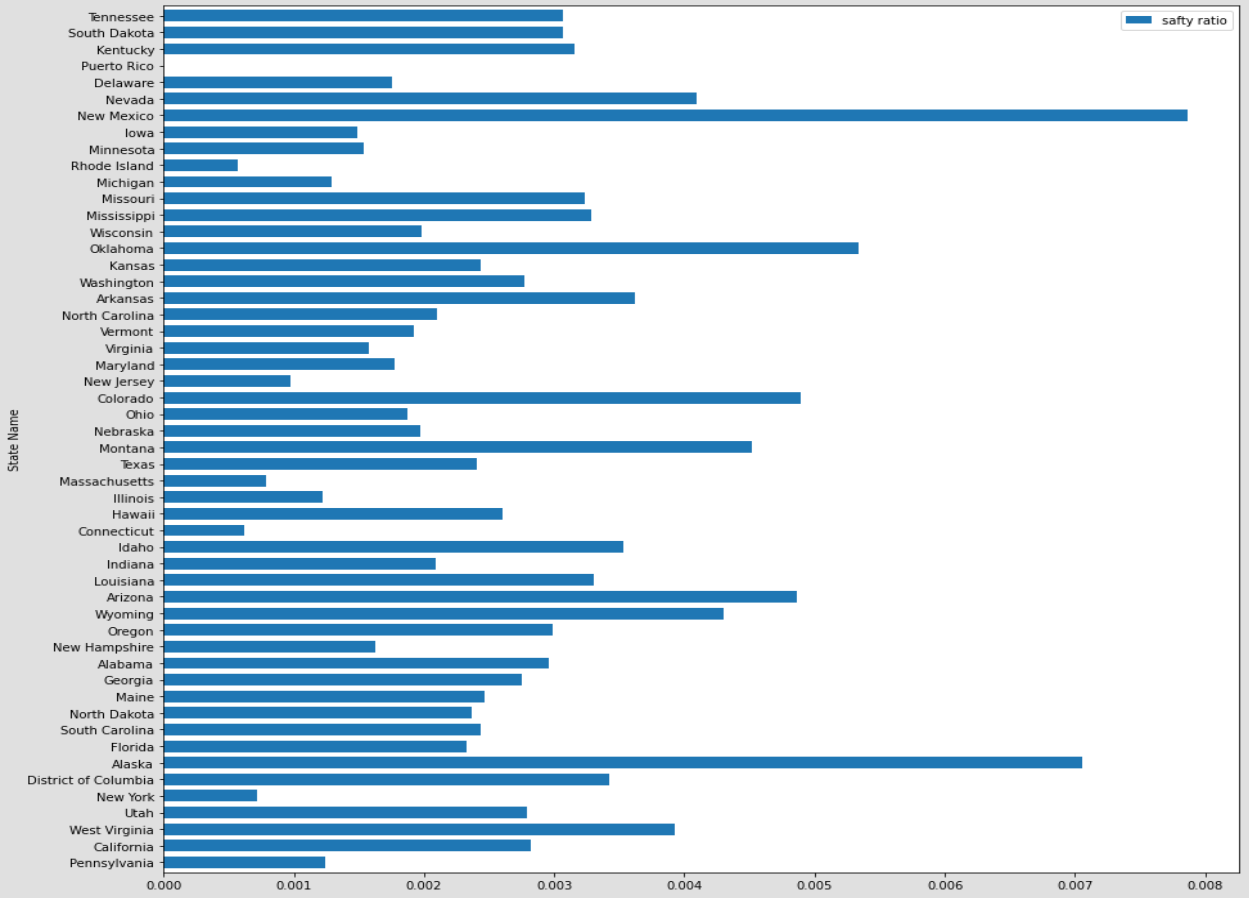
From the above plot it's clear that most of the states have white people more in number.

	percentage
Native	0.816891
Asian	5.587060
White	59.511292
Black	12.611323
Hispanic	18.971774
Others	2.501659

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:

```
(unarmNOTflee.at['B','counts']/census_race_pop.at['Black',0])*100000
```

```
0.15386357445758883
```

```
(unarmNOTflee.at['W','counts']/census_race_pop.at['White',0])*100000
```

```
0.052984696745909186
```

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:

```
(unarmflee.at['B','counts']/census_race_pop.at['Black',0])*100000
```

```
0.1899253497210862
```

```
2](unarmflee.at['W','counts']/census_race_pop.at['White',0])*100000
```

```
0.038210117845607586
```

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

Person fleeing armed:

```
](armflee.at['B','counts']/census_race_pop.at['Black',0])*100000
```

```
1.6900952006825773
```

```
(armflee.at['W','counts']/census_race_pop.at['White',0])*100000
```

```
0.5726422994461724
```

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:

```
)](armNOTflee.at['B','counts']/census_race_pop.at['Black',0])*100000
```

```
2.000226467948655
```

```
)](armNOTflee.at['W','counts']/census_race_pop.at['White',0])*100000
```

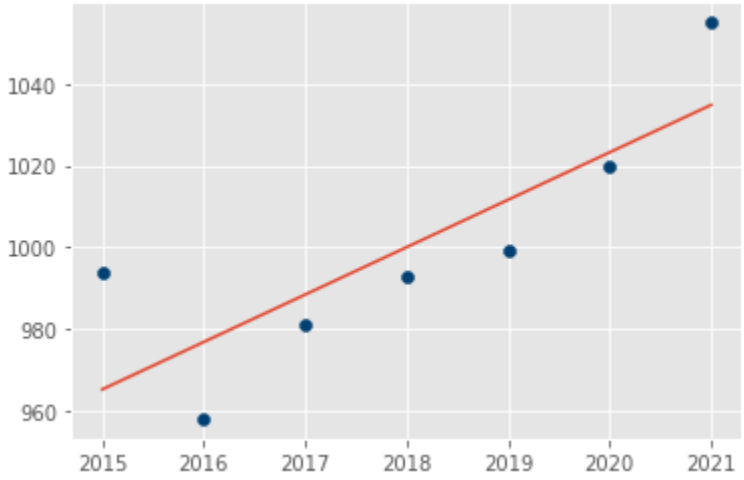
```
0.9578002873298969
```

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.

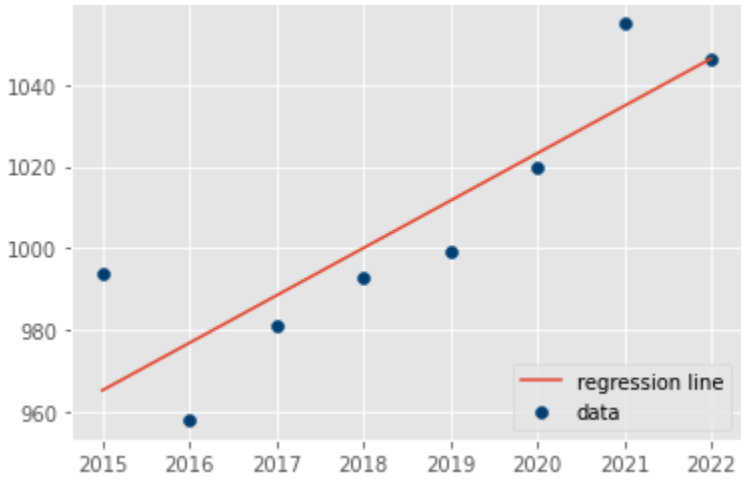
Linear Regression:

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.*'



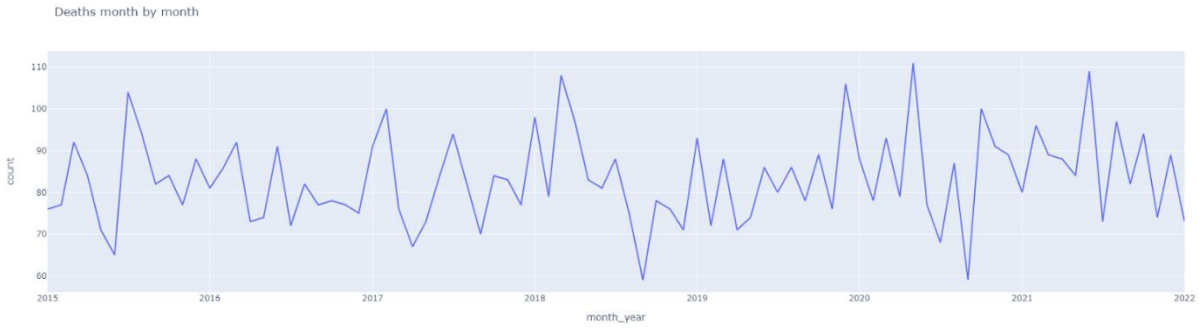
The predicted value for the year 2022 is 1047.

Auto Regressive Integrated Moving Average (ARIMA):

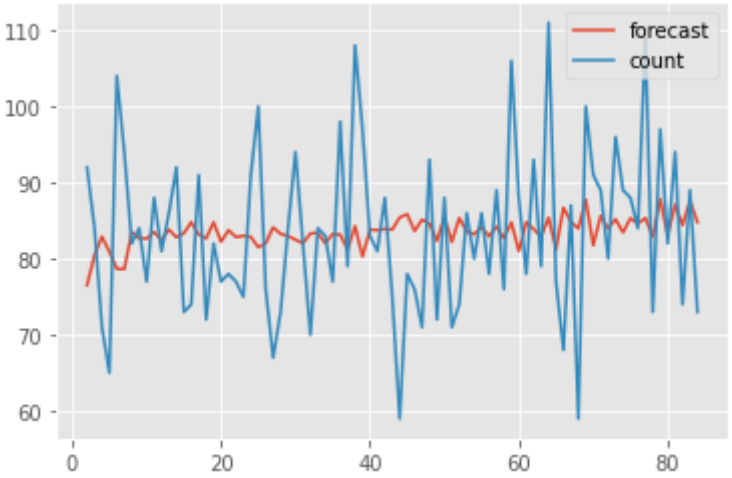
Auto Regressive Integrated Moving Average, is a set of models that explains a time series using its own previous values given by the lags (Auto Regressive) and lagged errors (Moving Average) while considering stationarity corrected by differencing (opposite of Integration.) In other words, ARIMA assumes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

AR: Autoregression.
I: Integrated.
MA: Moving Average.

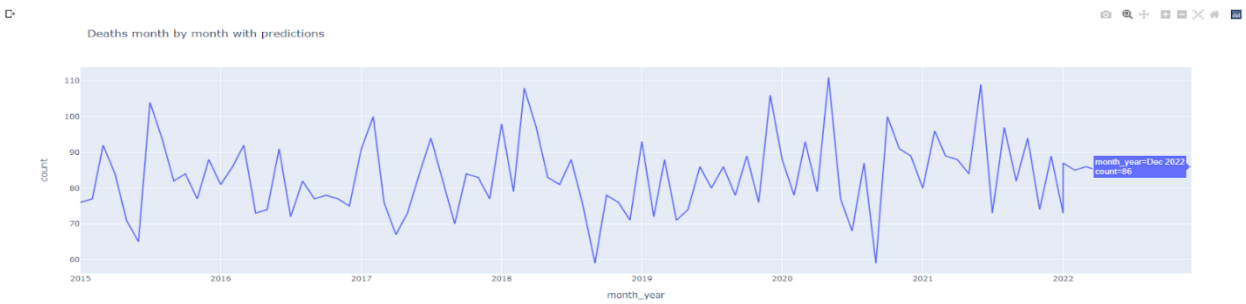
A standard notation is used of ARIMA(p,d,q).
p: Lag order.
d: Also called the degree of differencing (The number of times that the raw observations are different).
q: Also called the order of moving average (The size of the moving average window).
For our dataset we used (3,1,1) as this combination gave 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.

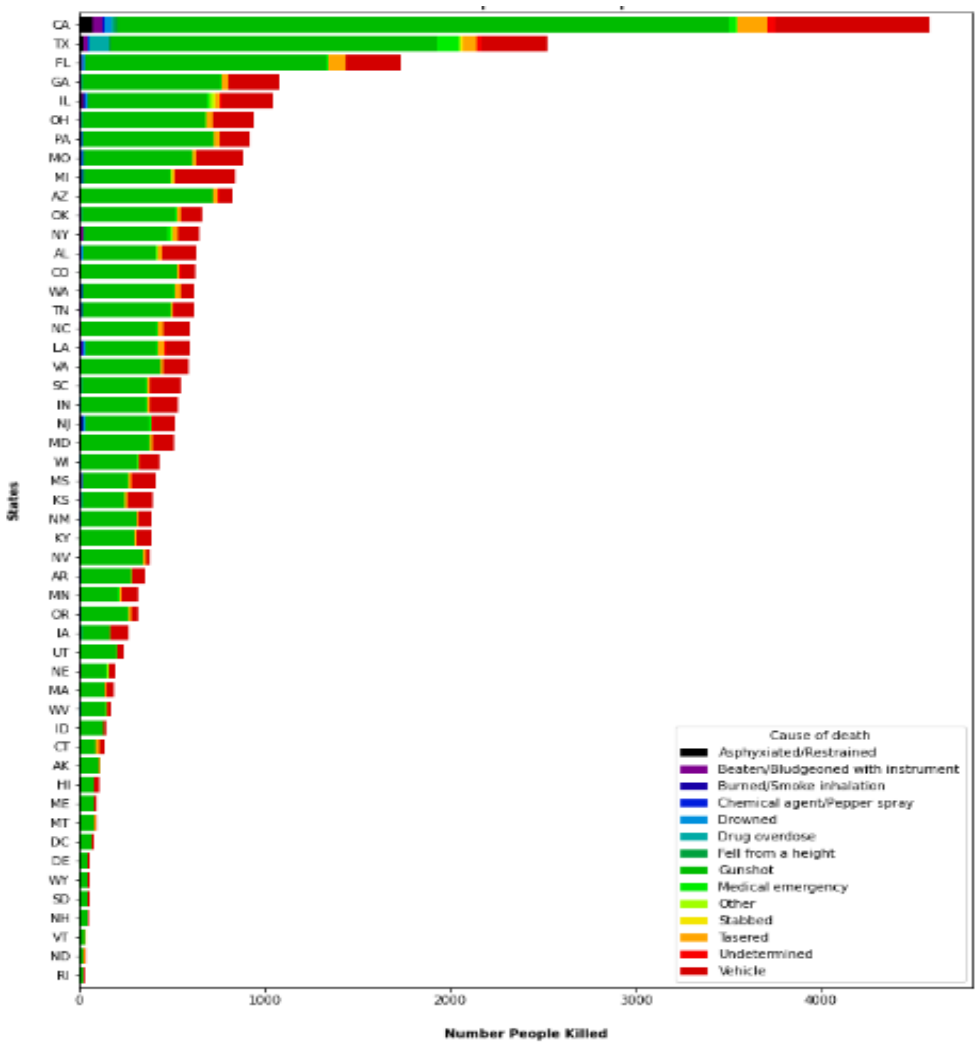
```
 #   Column                                                    Non-Null Count  Dtype
---  ------                                                    --------------  -----
 0   Unique ID                                                 28621 non-null  float64
 1   Subject's name                                            28622 non-null  object
 2   Subject's age                                             27608 non-null  object
 3   Subject's gender                                          28521 non-null  object
 4   Subject's race                                            28621 non-null  object
 5   Subject's race with imputations                           28448 non-null  object
 6   Imputation probability                                    28439 non-null  object
 7   URL of image of deceased                                  13130 non-null  object
 8   Date of injury resulting in death (month/day/year)        28622 non-null  object
 9   Location of injury (address)                              28080 non-null  object
 10  Location of death (city)                                  28586 non-null  object
 11  Location of death (state)                                 28621 non-null  object
 12  Location of death (zip code)                              28432 non-null  float64
 13  Location of death (county)                                28605 non-null  object
 14  Full Address                                              28621 non-null  object
 15  Latitude                                                  28621 non-null  float64
 16  Longitude                                                 28621 non-null  float64
 17  Agency responsible for death                              28553 non-null  object
 18  Cause of death                                            28621 non-null  object
 19  A brief description of the circumstances surrounding the death  28621 non-null  object
 20  Dispositions/Exclusions INTERNAL USE, NOT FOR ANALYSIS    28621 non-null  object
 21  Intentional Use of Force (Developing)                     28621 non-null  object
 22  Link to news article or photo of official document        28620 non-null  object
 23  Symptoms of mental illness? INTERNAL USE, NOT FOR ANALYSIS  28560 non-null  object
 24  Video                                                     9 non-null      object
 25  Date&Description                                          28587 non-null  object
 26  Unique ID formula                                         2 non-null      float64
 27  Unique identifier (redundant)                             28621 non-null  float64
 28  Date (Year)                                               28622 non-null  int64
dtypes: float64(6), int64(1), object(22)
```

The describe method returns the following output for our dataset.

	Unique ID	Location of death (zip code)	Latitude	Longitude	Unique ID formula	Unique identifier (redundant)	Date (Year)
count	28621.000000	28432.000000	28621.000000	28621.000000	2.000000	28621.000000	28622.000000
mean	14311.000000	58432.535488	36.767127	-95.441638	28059.000000	14311.000000	2011.190972
std	8262.315364	27982.930895	5.160217	16.339723	794.788022	8262.315364	5.837947
min	1.000000	1013.000000	19.034681	-165.591880	27497.000000	1.000000	2000.000000
25%	7156.000000	33159.250000	33.542220	-111.278099	27778.000000	7156.000000	2006.000000
50%	14311.000000	60660.500000	36.692833	-90.556579	28059.000000	14311.000000	2012.000000
75%	21466.000000	85044.000000	40.426677	-82.576535	28340.000000	21466.000000	2016.000000
max	28621.000000	99921.000000	71.301250	-67.266033	28621.000000	28621.000000	2100.000000

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.



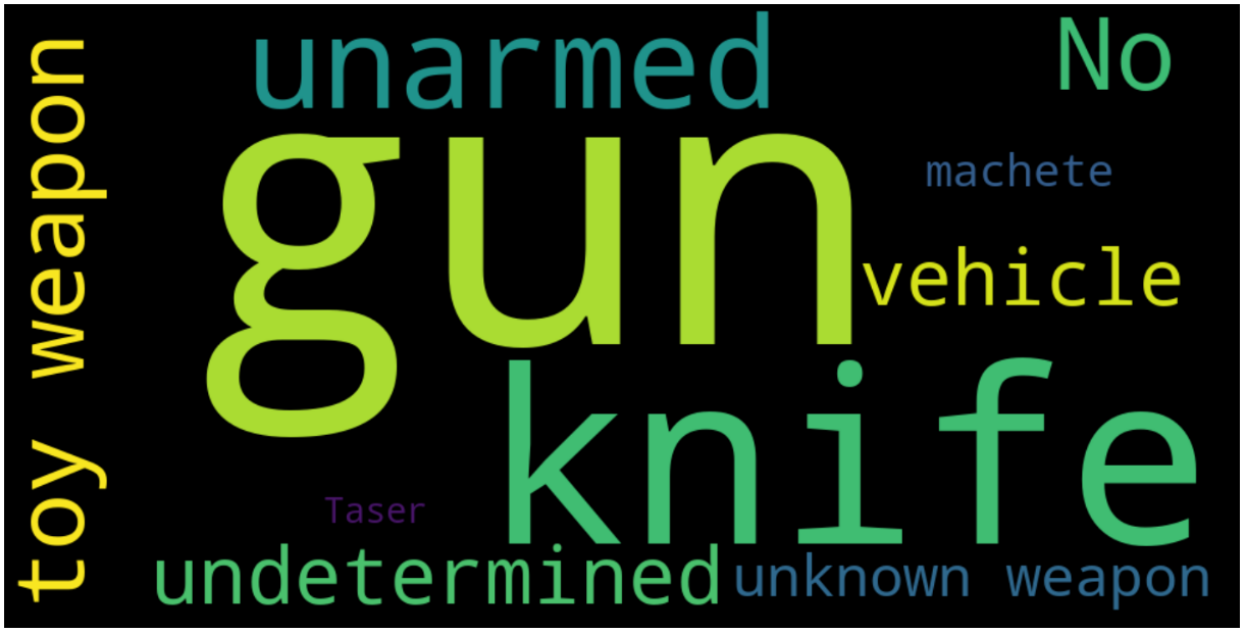
Causes of People deaths per City:

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.

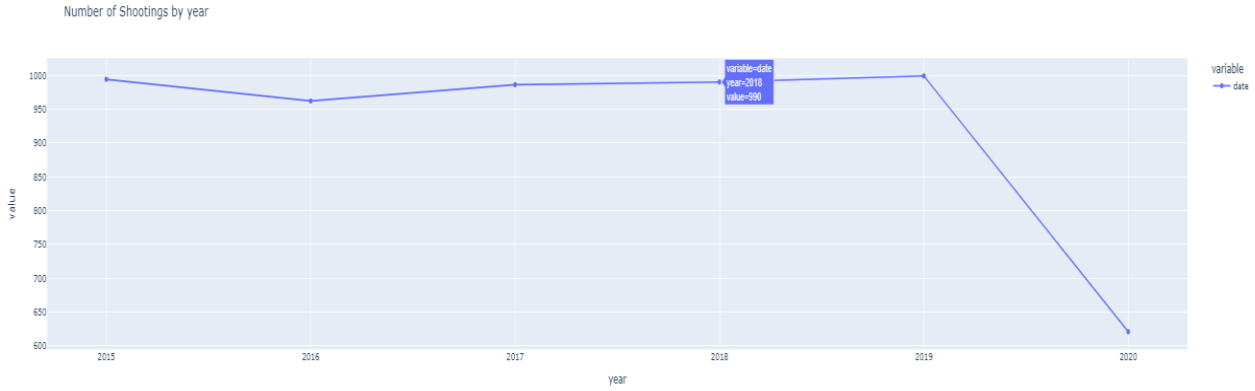
After that, we have used several plots to represent the data by year, month and day to know the number of killings as per it respectively.

	id	name	date	manner_of_death	armed	age	gender	race
0	3	Tim Elliot	2015-01-02	shot	gun	53.0	Male	Asian
1	4	Lewis Lee Lembke	2015-01-02	shot	gun	47.0	Male	White
2	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23.0	Male	Hispanic
3	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32.0	Male	White
4	9	Michael Rodriguez	2015-01-04	shot	nail gun	39.0	Male	Hispanic

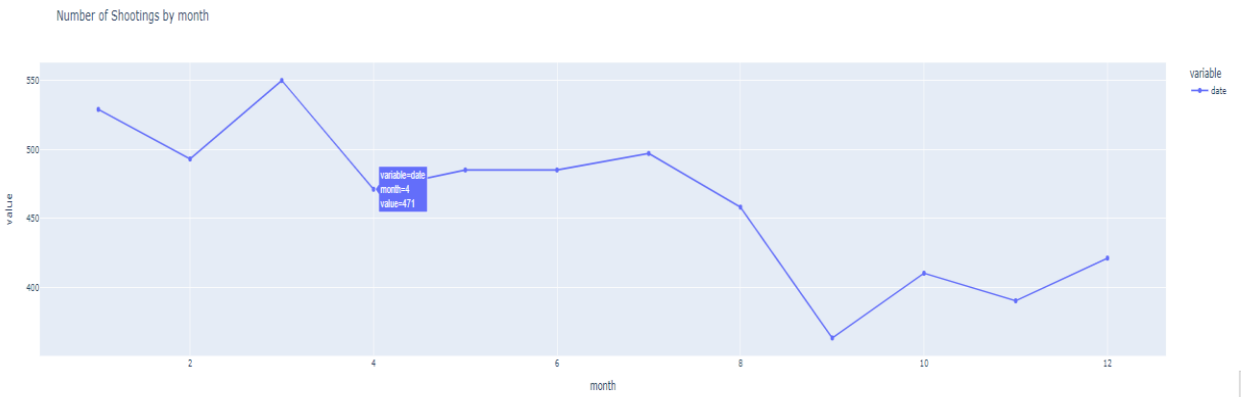
The above shown is the bar graph represented for the killings from the year 2015 to 2020 taking the count on y axis and year on x axis. Among them 2015 and 2019 years has the most number of shootings.

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

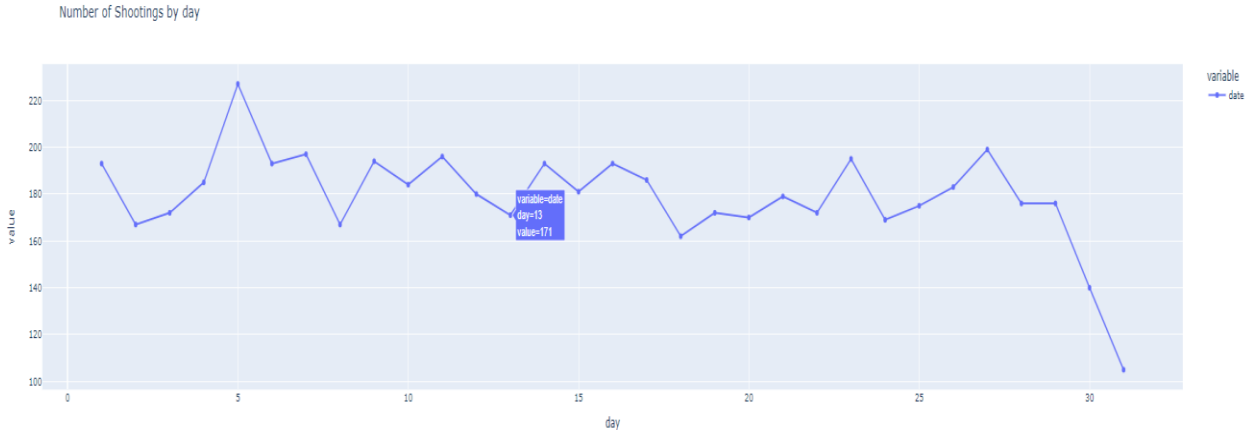
1. Number of shootings per year:



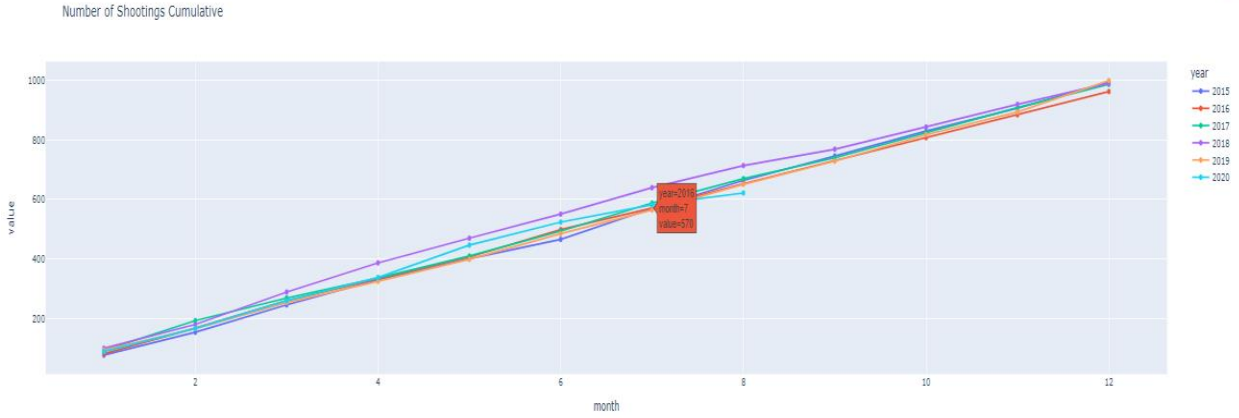
2. Number of shootings per month



3. Number of shootings per day



4. Number of shootings cumulative



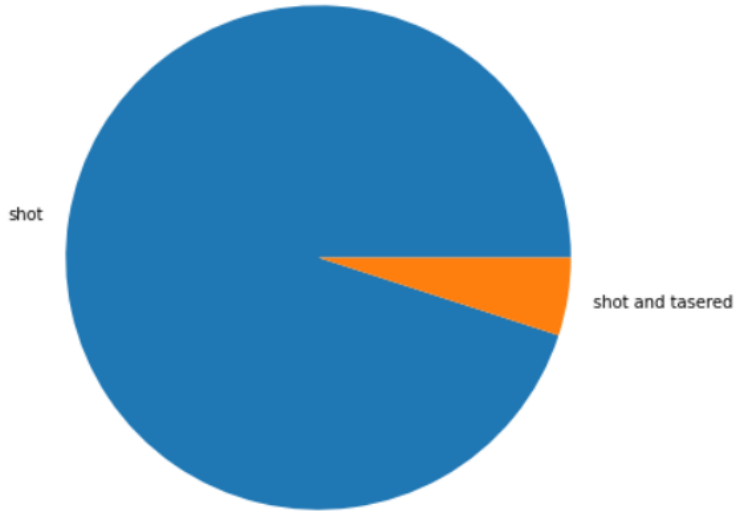
Analysis of Pie Charts for Manner of Death:

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.

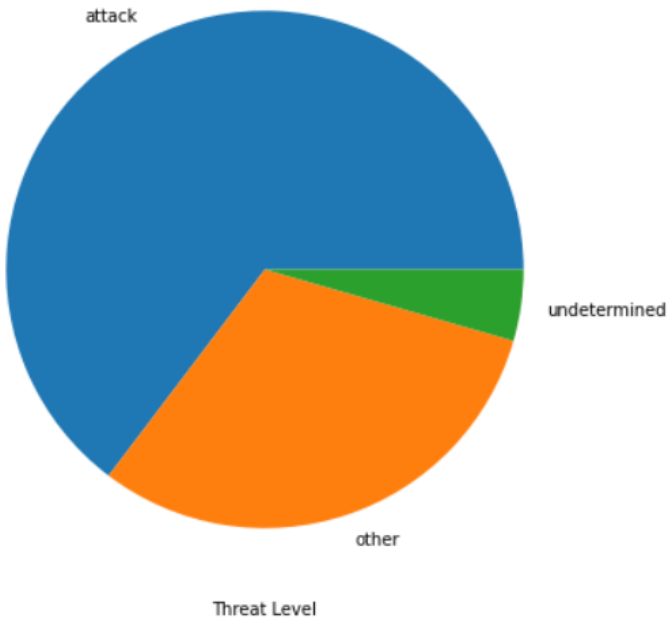
Using the labels ‘attack’, ‘other’, ‘undetermined’ we have represented a pie chart for number of shootings for these manner of deaths. From that we got to know that 3591 number of people have been attacked whereas 1714 number of people are of other category and tasered and 247 have been undetermined.

1. Using the labels shot and shot and tasered:




```
shot                5275
shot and Tasered     277
Name: manner_of_death, dtype: int64
```

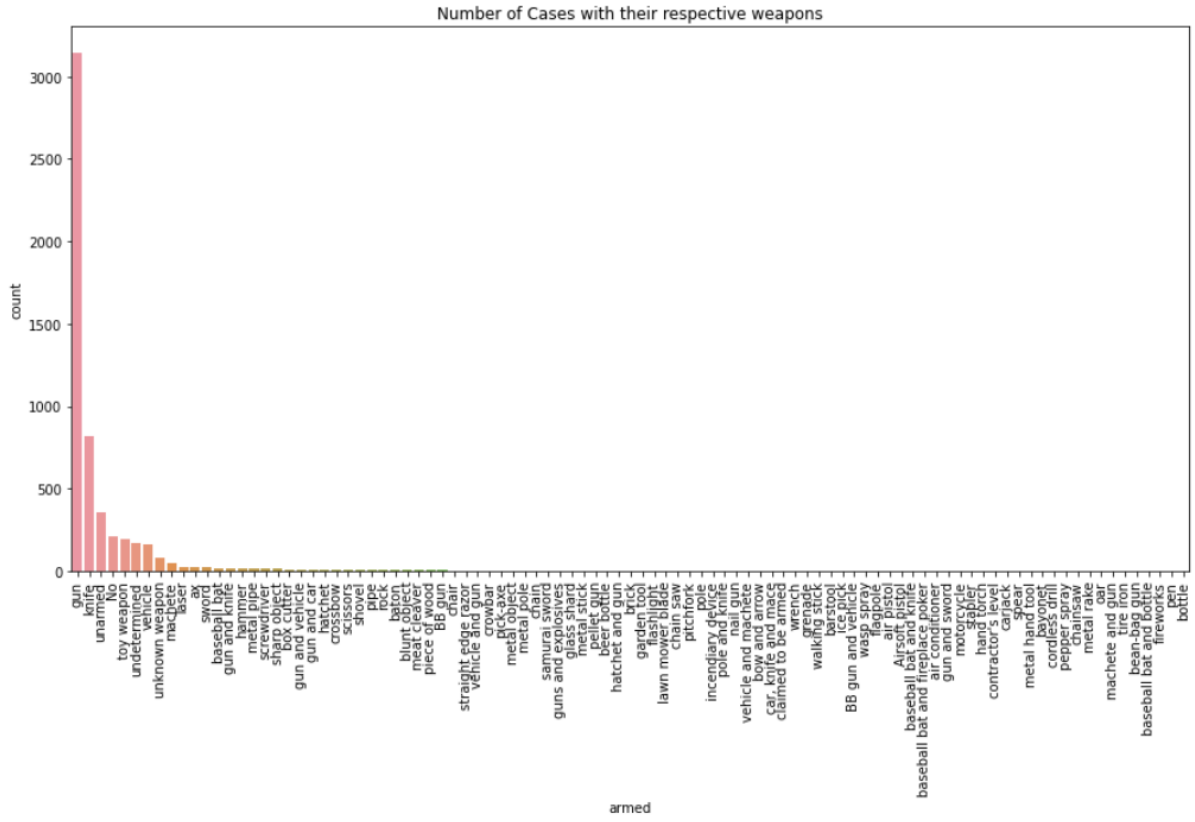
2. Using labels attack, undetermined and other:




```
attack          3591
other           1714
undetermined     247
Name: threat_level, dtype: int64
```

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.

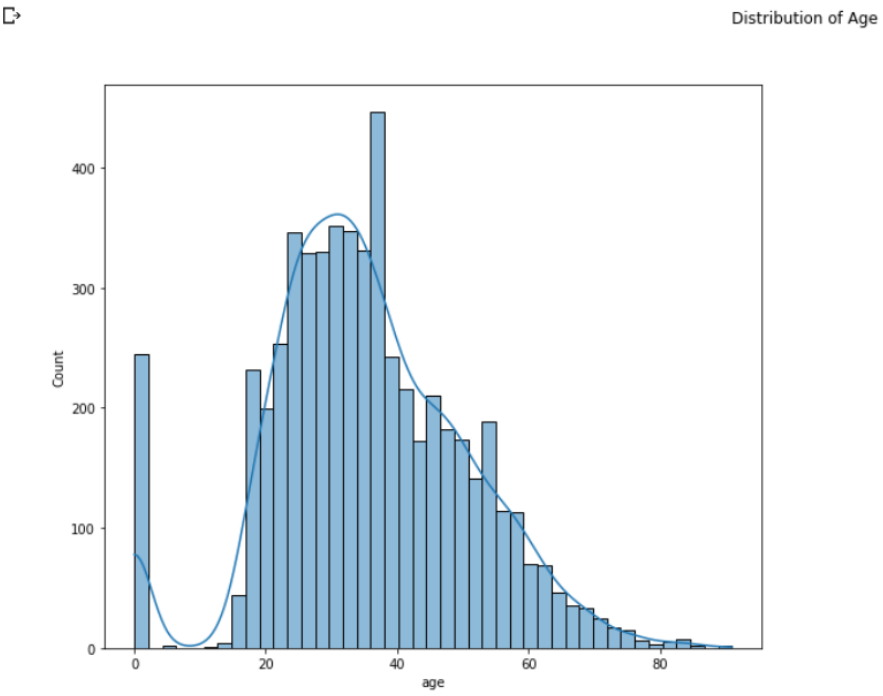


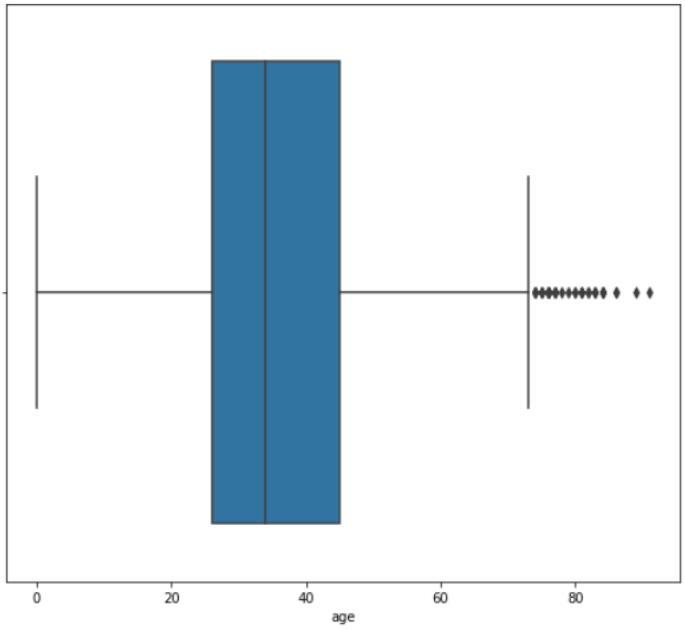
gun	3146
knife	818
unarmed	355
No	213
toy weapon	193
undetermined	168
vehicle	159
unknown weapon	79
machete	46
Taser	26
ax	24
sword	23
baseball bat	18
gun and knife	18
hammer	16
metal pipe	13
screwdriver	13
sharp object	13
box cutter	12
gun and vehicle	11
gun and car	11
hatchet	11
crossbow	9
scissors	7
shovel	6
pipe	6

Name: armed, dtype: int64

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.

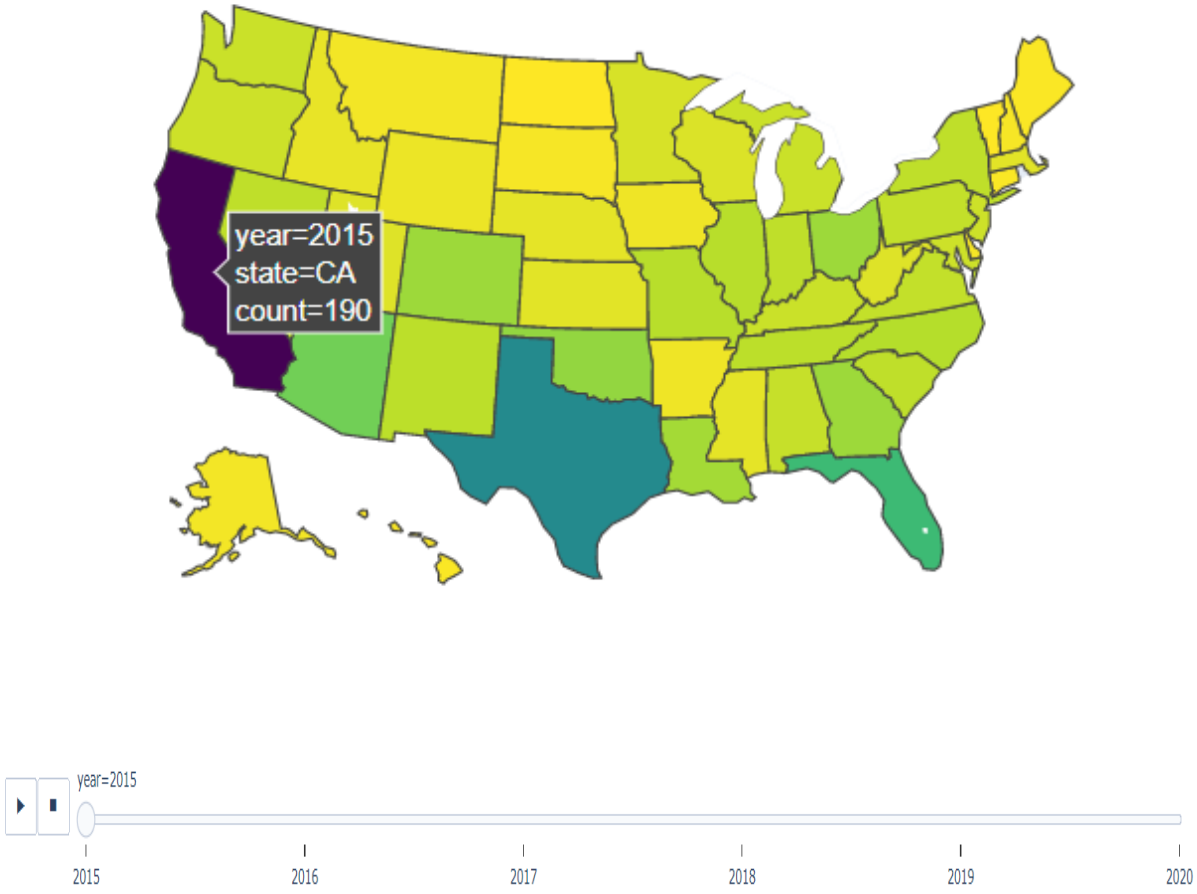




Most of the people's age who were involved in the police shootings lie between 20-45 as most of the people were young.

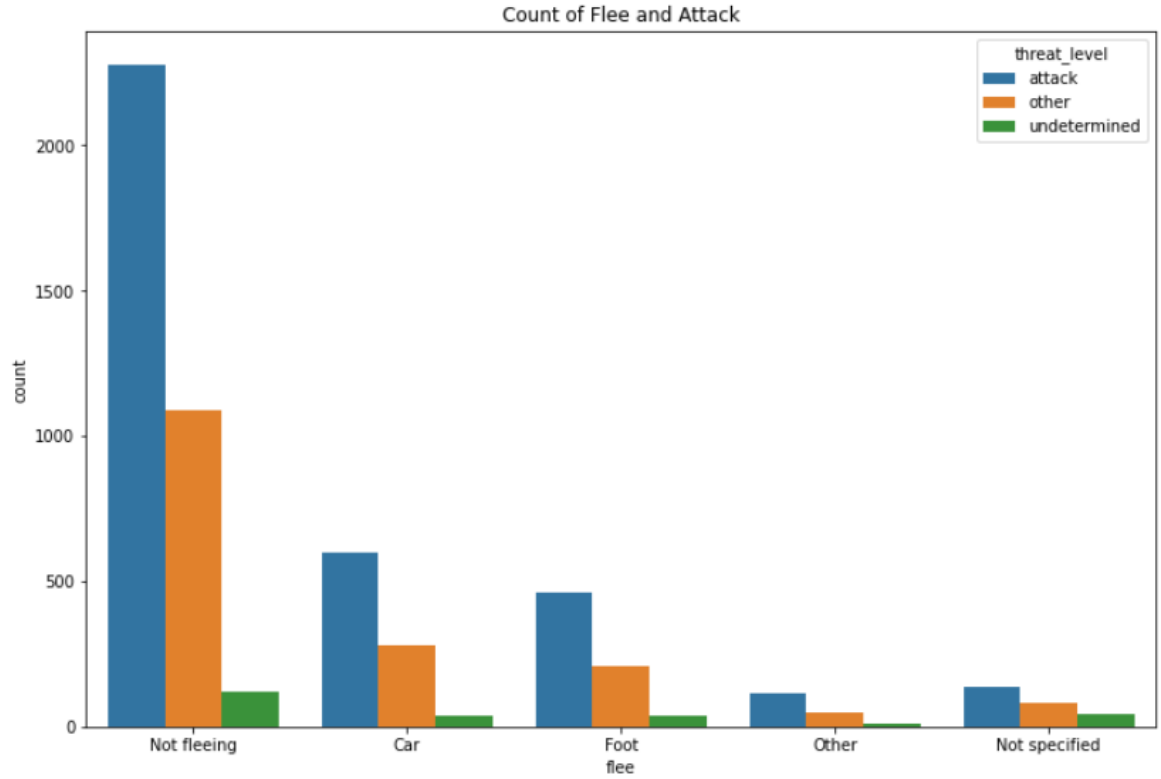
Incidents Observed in Each State Over the Year:

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:




```
Not fleeing      3489
Car               917
Foot              710
Not specified     262
Other             174
Name: flee, dtype: int64
```

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

Conclusion:

From this project, we got to understand the current situation of many races, the work that police men have to take up, prepare themselves in day to day life.

On the other hand, the situation of African American race people has to face a lot of discrimination on a daily basis. The chances of them getting killed for the same crime the other races do is more.

While doing this project, we learnt how to perform various visualizations, linear and time forecasting for predictions.

Using APIs to get data that we need. We got hands-on experience on a real time project for data mining that's going to help us boost our morale to work on many more such projects in the future.

References:

https://dataindependent.com/pandas/pandas-bar-plot-dataframe-plot-bar/#:~:text=Pandas%20Bar%20Plot%20is%20a,a%20chart%20for%20you%20automatically

https://phdinds-aim.github.io/time_series_handbook/01_AutoRegressiveIntegratedMovingAverage/01_AutoRegressiveIntegratedMovingAverage.html

https://realpython.com/linear-regression-in-python/

https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

https://www.statista.com/statistics/585152/people-shot-to-death-by-us-police-by-race/