



**Spring 2023**

**CMPE 255 - DATA MINING**

**PROJECT REPORT**

**On**

**“SHOOTING ANALYSIS IN U.S.A”**

**Instructor:**

**TAEHEE JEONG**

**Submitted By:**

**SRIKARI VEERUBHOTLA [015957032]**

**PRANITHA HARANI KOKA [016676491]**

**PRATIKSHA RAMARAO MASALKAR [016711929]**

**BHARADWAJ ROUTHU [016597932]**

## **CONTENTS:**

- I. Introduction
- II. Literature Review/ Background Study
- III. Data Collection and Preprocessing
- IV. Data Visualization
- V. Methodology OR Implementation
- VI. Results and Discussion
- VII. Data Analysis
- VIII. Individual Contributions
- IX. Conclusion
- X. Contribution to the Society
- XI. References

## **Abstract:**

This project investigates the prevalence and causes of shootings in the United States. Using data from various sources including news reports, government statistics, and academic research, we analyzed trends in the frequency and characteristics of shootings across different regions and demographic groups. We found that shootings are a significant public health and safety issue in the US, with a high number of fatalities and injuries each year. Our analysis revealed that shootings are disproportionately concentrated in certain regions and demographic groups, and that factors such as access to firearms, mental health issues, and social and economic inequality are significant drivers of the problem. Based on our findings, we provide recommendations for policy interventions to reduce the prevalence and impact of shootings in the US. This project contributes to a better understanding of the nature and scope of the problem of shootings in the US and provides insights that can inform effective policy responses.

## **I. Introduction**

### **A. Overview of the Project**

In recent years, there have been many incidents of shootings in the United States that have resulted in injury or death. These shootings have raised questions about the causes and consequences of shootings, and there is much we do not know about these issues. In this report, we aim to provide a detailed analysis of shootings in the US, using a variety of data sources and research methods. We want to find out how often shootings occur, who is most likely to be involved, and what factors may contribute to these incidents.

Shootings in the United States are a persistent and devastating problem. Whether it is mass shootings, homicides, or suicides, gun violence has become a public health crisis that impacts individuals, families, and communities across the country. Despite the widespread concern about shootings, there is still much we do not know about the nature and extent of this problem.

### **B. Purpose and objectives of the project**

The purpose of this project is to conduct a thorough analysis of shooting data in the United States. Specifically, we aim to study six shooting datasets which are

```
fatal_encounters_dot_org.csv, police_killings.csv,  
deaths_arrests_race.csv, police_deaths_538.csv,  
shootings_wash_post.csv, shootings.csv
```

- Examine factors such as location, time, type of shooting, and other relevant variables

- Identify any patterns or trends that may exist in the data
- Provide insight into the nature and extent of shootings in the country
- Contribute to the ongoing discussion and debate surrounding shootings in the United States
- Identify potential avenues for policy and reform in this area.

### **C. Scope of the project**

This project report aims to provide a comprehensive analysis of shootings in the US, drawing on a range of data sources and analytical methods. We will examine the frequency, distribution, and demographic characteristics of shootings, as well as the contextual factors and individual-level predictors that may contribute to these incidents.

## **II. Literature Review/ Background Study**

Several studies have highlighted the role of firearms in facilitating shootings.

For example, one study found that states with higher levels of gun ownership had a higher incidence of firearm fatalities, including homicides and suicides (Miller, Azrael, & Hemenway, 2007).

Another study found that stricter gun control policies were associated with lower rates of gun-related deaths (Kalesan, Mobily, & Keiser, 2016). These findings suggest that reducing access to firearms may be an effective strategy for reducing the incidence of shootings.

Other research has focused on individual-level predictors of shootings. Several studies have found that individuals with a history of mental illness are more likely to engage in shootings (Swanson et al., 2015; Stone et al., 2017). Other studies have highlighted the role of social and environmental factors, such as exposure to violence and poverty, in contributing to shootings (Sampson, Raudenbush, & Earls, 1997; Sampson & Wilson, 1995).

One study conducted by researchers at UCLA's Department of Psychology examined racial disparities in police shootings (Plant, Peruche, & Butz, 2005). The study found that Black suspects were more likely to be shot by police than White suspects, even when controlling for factors such as whether the suspect was armed or not.

Another study conducted by researchers at UCLA's Luskin School of Public Affairs looked at the relationship between community policing and police shootings (Bailey & Linski, 2017). The study found that police departments that engaged in community policing practices had lower rates of police shootings than departments that did not.

### Limitations and challenges of data mining:

1. **Incomplete or biased data:** Data on shootings in the US may not be comprehensive or may be influenced by selection bias, which can lead to inaccurate results.
2. **Difficulty in obtaining data:** Data on shootings may not be easily accessible or may not be available in a usable format for analysis.
3. **Data quality issues:** The accuracy and completeness of the data may be questionable, which can lead to incorrect insights.
4. **Interpretation of results:** The results of data mining can be complex and may require specialized knowledge to interpret, which can be a challenge for non-experts.
5. **Privacy concerns:** Collecting and analyzing data on individuals involved in shootings can raise privacy concerns and ethical issues.
6. **Technical challenges:** The analysis of large and complex datasets can be computationally intensive and require specialized skills and resources.
7. **Limited variables:** The available data may not contain all the relevant variables, which can limit the scope of analysis and the accuracy of results.

### III. Data Collection and Preprocessing

#### A. Data sources and types

We obtained our data from various sources. In total, we used four datasets for our analysis, with three datasets acquired from Kaggle and one from the Washington Post. The types of data included in these datasets varied, ranging from demographic information such as age, gender, and race, to location and date information related to the incidents. The datasets provided information on various types of shootings, such as homicides, suicides, and accidents. By utilizing multiple datasets, we aimed to obtain a comprehensive and diverse representation of shooting incidents in the United States.

Dataset 1 columns and their description. ( US Police shooting in from 2015-22 )

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	Were 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 ?

Dataset 2 variables and their descriptions. (police\_deaths\_538.csv)

<b>Column Names</b>	<b>Description</b>
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.

## **B. Data cleaning and preprocessing techniques**

The following are the steps taken in data preprocessing:

1. To identify missing data, we calculated the percentage of null values in each column. The Race column had the highest number of null values.
2. Missing values were found in the name field. Instead of randomly assigning names, we decided to remove the column altogether as it would not be useful in the model.
3. In the Race column, we chose to categorize values as "Other" because solely considering race may not provide accurate information.
4. We corrected the age column by ensuring that all values were between 0 and 100. Some values were negative, which could impact the accuracy of predictions based on the dataset.
5. The date entry was checked for validity and converted to a specified format of MM-DD-YYYY.

## **C. Data transformation and normalization**

Data transformation involves converting the data from its original form into a new format that is more suitable for analysis.

Normalization is the process of adjusting the values of a dataset to have a common scale or distribution.


## **D. Data integration and feature selection**

Data integration involves combining multiple datasets into a single dataset, while feature selection involves identifying and selecting the most relevant variables or features for analysis.



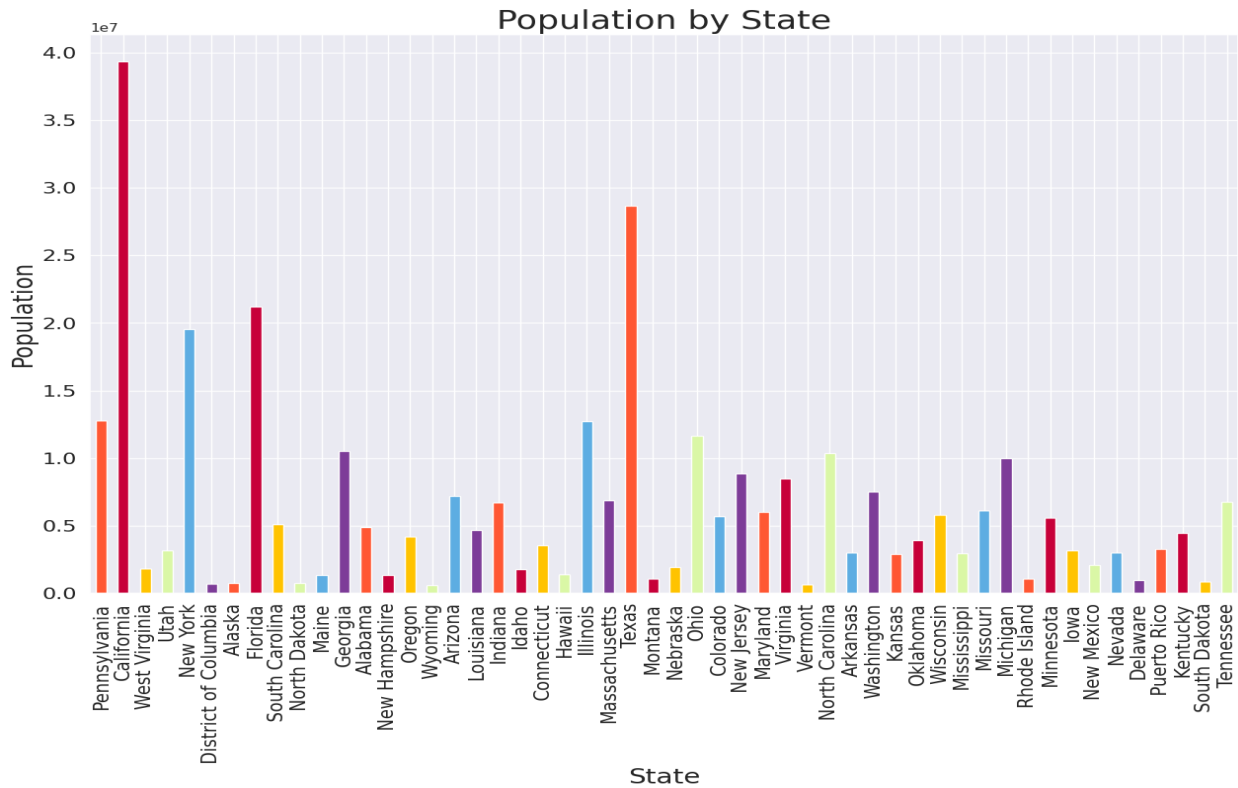
IV. Data Visualization

The graph displays the total population of each state in the United States as sourced from the US Census Bureau's American Community Survey (ACS) for 2020. The y-axis represents the total population count, while the x-axis represents each state. The bars are colored according to a predefined color palette and the graph is useful for visually comparing the population of each state and identifying any outliers.



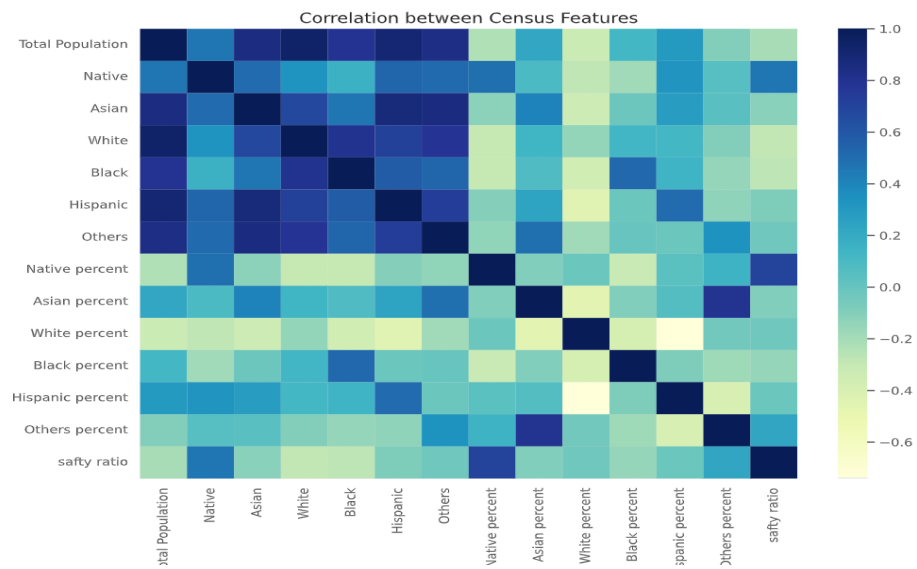
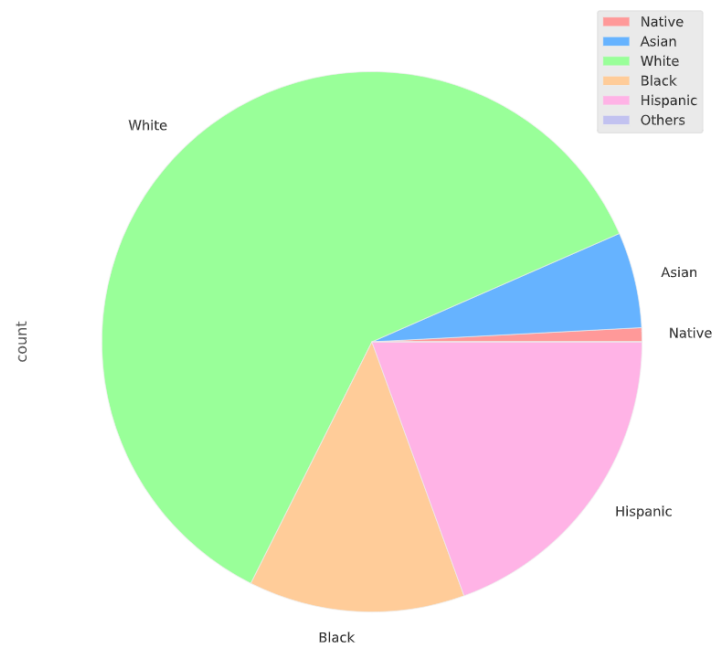
	Total Population	Native	Asian	White	Black	Hispanic	state
State Name							
Wyoming	581348	13117	4892	485816	5079	58854	56
Vermont	624340	1873	10126	576601	7964	12518	50
District of Columbia	701974	2438	28762	257792	318631	77981	11
Alaska	736990	107298	47289	439979	23894	53059	02
North Dakota	760394	39165	11979	636284	23959	30325	38
South Dakota	879336	74975	12413	715328	18836	36088	46
Delaware	967679	3560	38528	595236	212795	91350	10
Rhode Island	1057798	4344	36536	755708	69196	168007	44
Montana	1061705	65523	8664	908782	5919	41501	30
Maine	1340825	8894	15270	1242113	18635	23143	23
New Hampshire	1355244	2197	36581	1214029	21045	52792	33

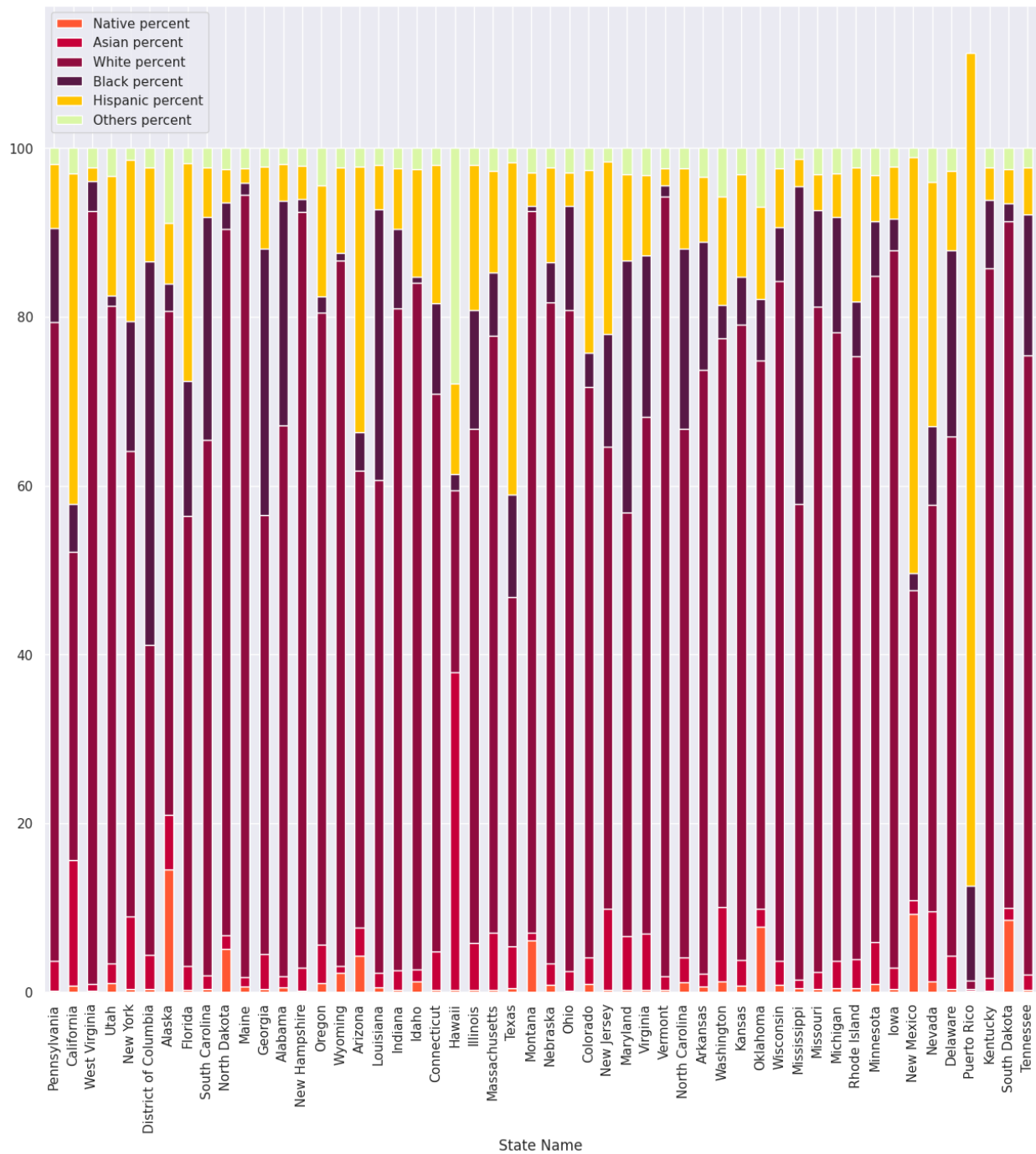
According to the visualization, the population of the United States is highly diverse, with a wide variety of nationalities and ethnic backgrounds. It also shows that certain states have a higher population of certain racial groups than others.



We can identify which states have the highest and lowest shooting incidents by plotting the number of shootings per state as a bar graph. Using this data to identify patterns and trends in the data and build predictive models based on the characteristics of the state can provide valuable insights for further analysis and modeling.

Visualization is a crucial step in data mining as it helps to uncover patterns and relationships that might otherwise be difficult to see in raw data. Using this graph, we can quickly identify states with the highest and lowest number of shootings. This helps us determine which states need more attention to policy and prevention.

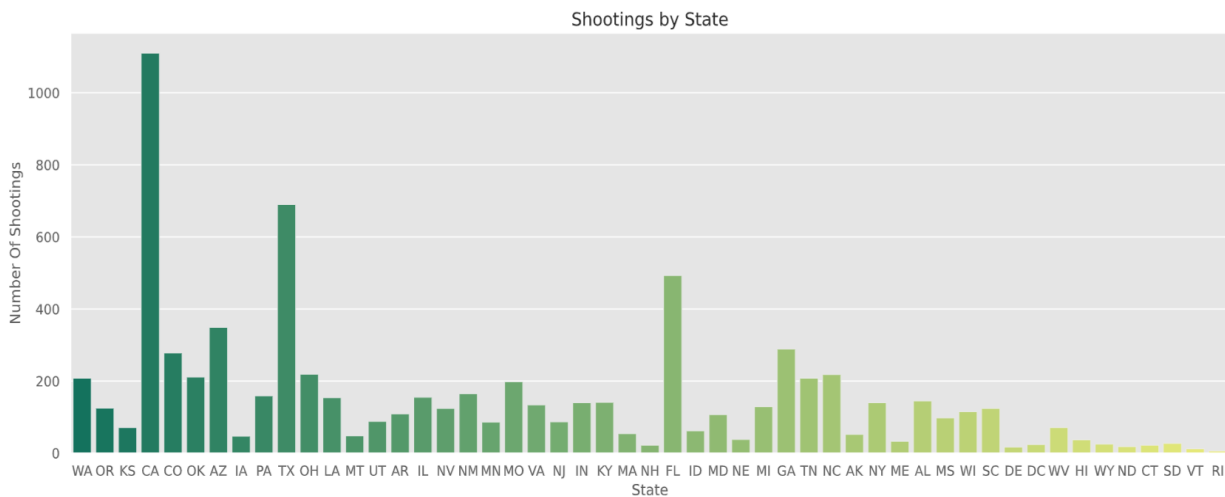




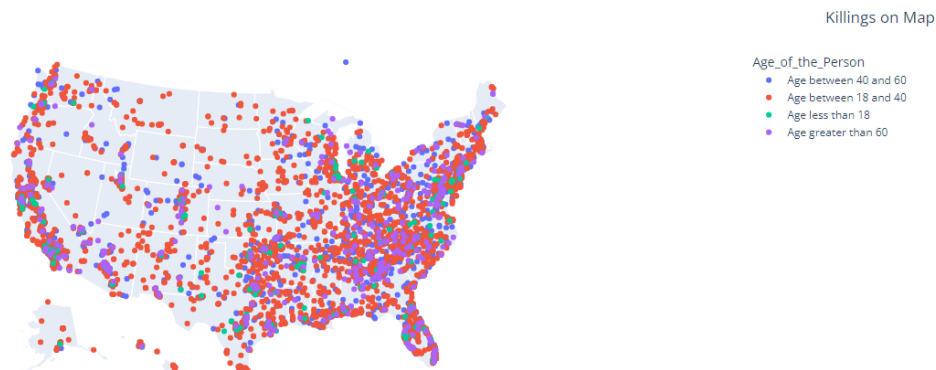
The positive correlation between two variables means that as one variable increases, the other also increases, whereas the negative correlation means that as one variable increases, the other decreases.

Correlation visualization is an important part of data mining since it helps to identify patterns and relationships between variables. For example, a predictive model can identify highly correlated features that may be redundant. Moreover, it can be used to identify features that have a strong correlation with the target variable, which is useful for building predictive models

**Coming to the shooting analysis :**



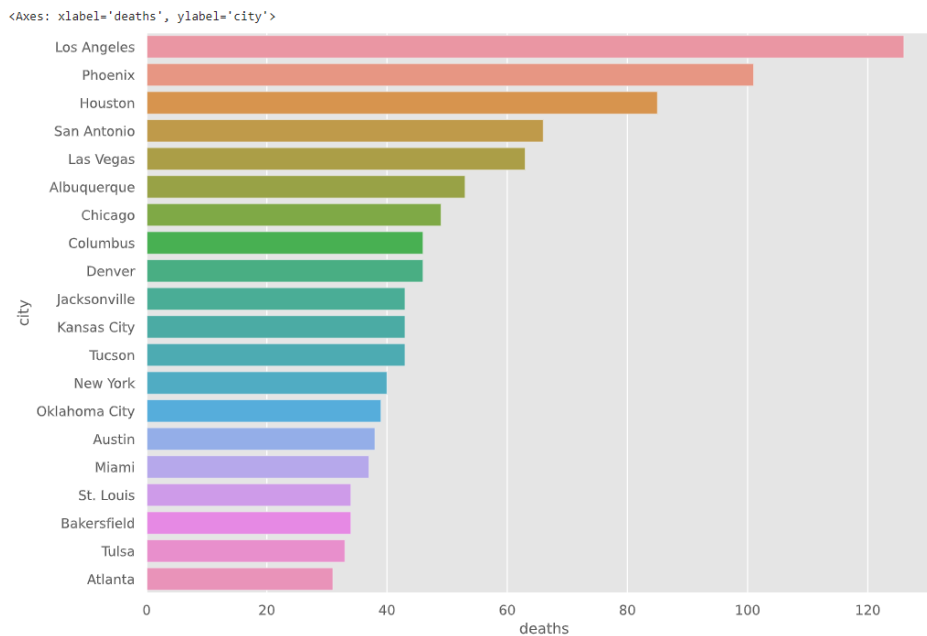
**For better visualization**



There is a graphic that summarizes and visualizes the causes of death in different states. By analyzing the data, trends and patterns can be identified and decision-making processes related to public health and safety can be informed.

California having more number of deaths we are trying to analysis in the big cities for better understanding of the data

## Causes of deaths by City:



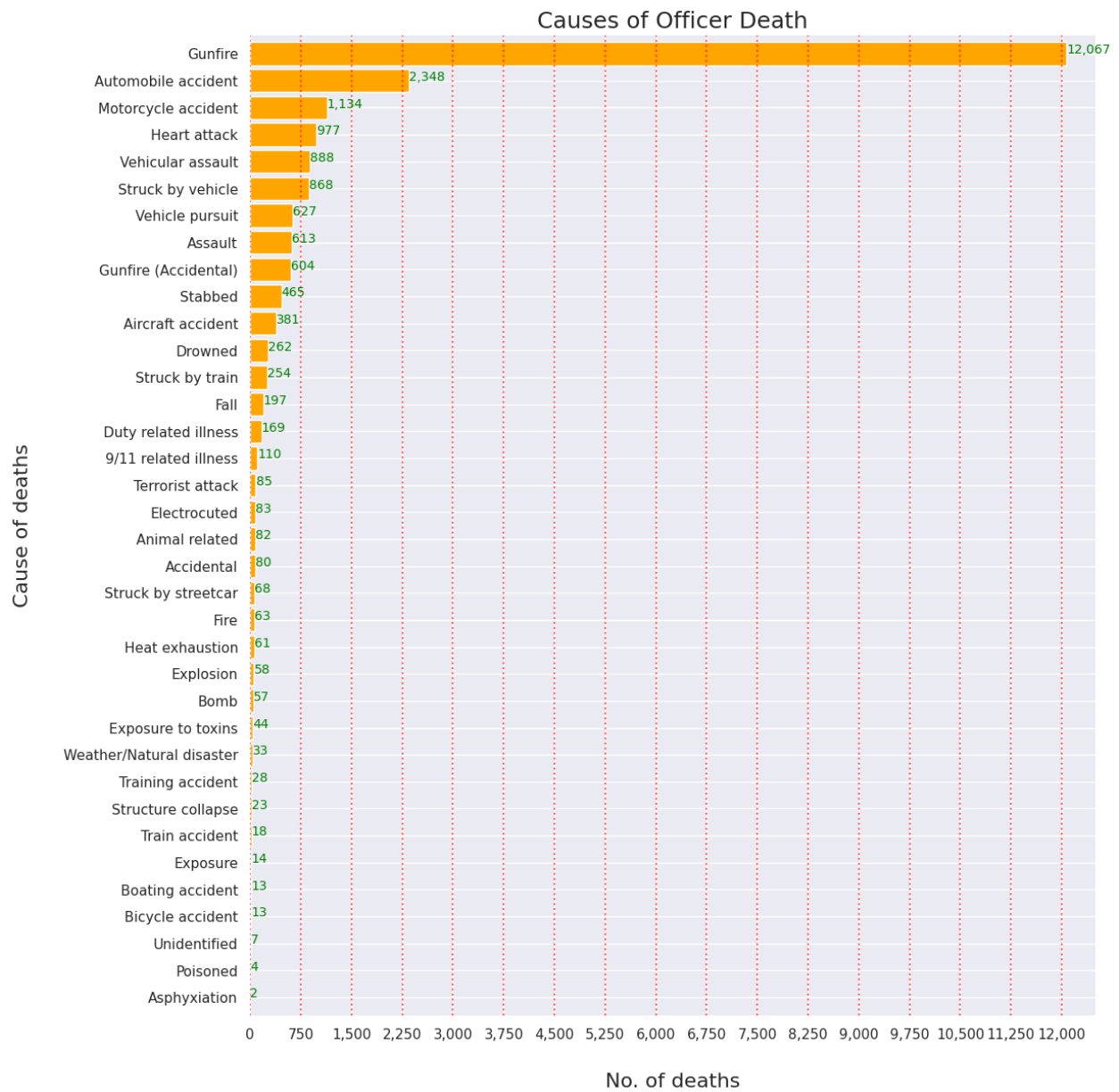
The deaths of citizens are not the only thing we have to analyze; we also need to analyze the offices involved to get the socioeconomic impact of the incident:

Using this analysis, we can gain valuable insight into the causes of officer deaths in the U.S. Knowing the risks officers face in the line of duty will enable law enforcement agencies to mitigate those risks. Additionally, it can provide guidance on officer training, equipment, and support policies. Using data visualization and pivot tables to gain insights from a dataset and present the results in an easy-to-understand format, this code illustrates data mining.

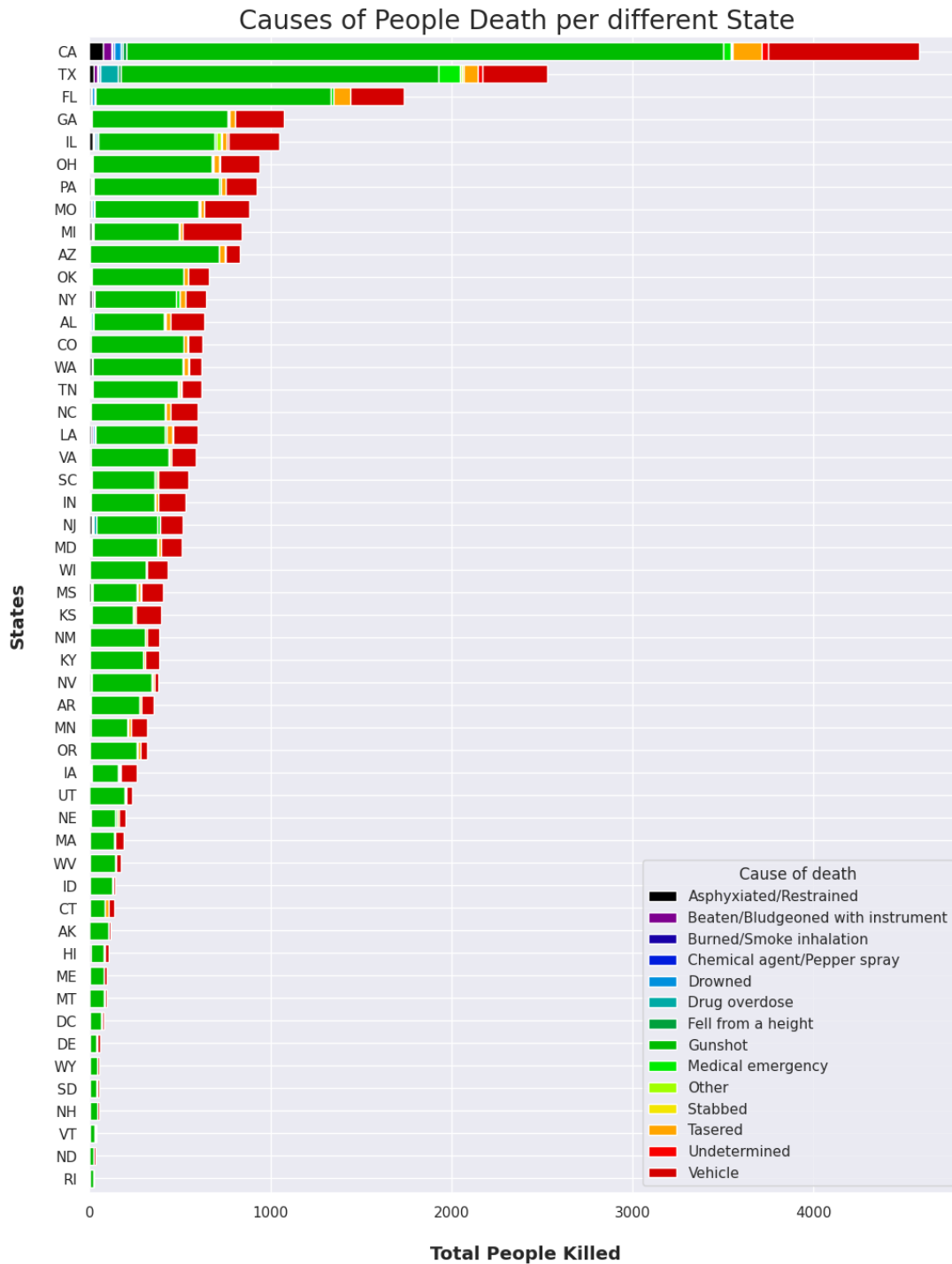
	person	dept	eow	cause	cause_short	date	year	canine	dept_name	state
0	Constable Darius Quimby	Albany County Constable's Office, NY	EOW: Monday, January 3, 1791	Cause of Death: Gunfire	Gunfire	1791-01-03	1791	False	Albany County Constable's Office	NY
1	Sheriff Cornelius Hogeboom	Columbia County Sheriff's Office, NY	EOW: Saturday, October 22, 1791	Cause of Death: Gunfire	Gunfire	1791-10-22	1791	False	Columbia County Sheriff's Office	NY
2	Deputy Sheriff Isaac Smith	Westchester County Sheriff's Department, NY	EOW: Thursday, May 17, 1792	Cause of Death: Gunfire	Gunfire	1792-05-17	1792	False	Westchester County Sheriff's Department	NY
3	Marshal Robert Forsyth	United States Department of Justice - United S...	EOW: Saturday, January 11, 1794	Cause of Death: Gunfire	Gunfire	1794-01-11	1794	False	United States Department of Justice - United S...	US
4	Sheriff Robert Maxwell	Greenville County Sheriff's Office, SC	EOW: Sunday, November 12, 1797	Cause of Death: Gunfire	Gunfire	1797-11-12	1797	False	Greenville County Sheriff's Office	SC

Visualization of the the above data

### Causes of officer death:



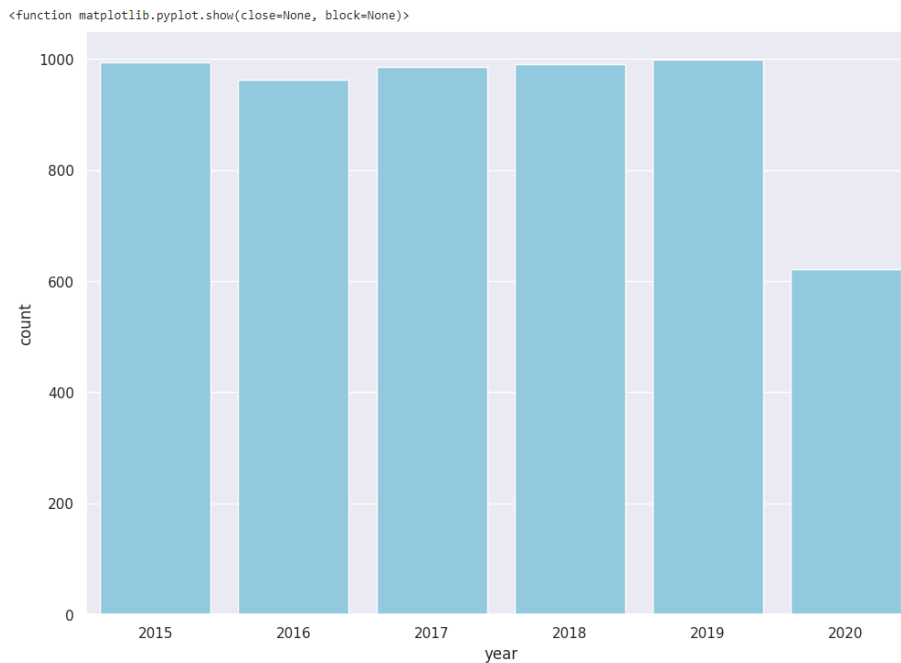
## Causes of people's death per state:



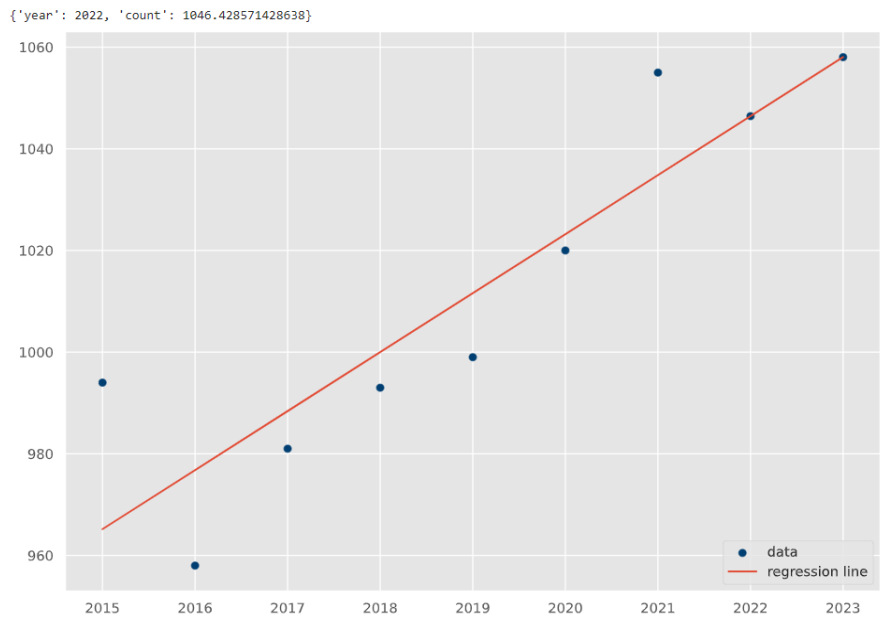
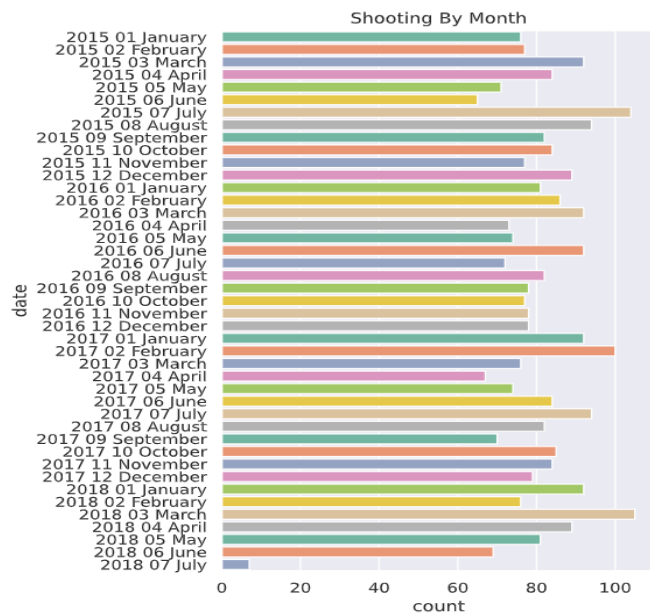
A bar chart showing the number of shootings from 2015 to 2020, using data from the "shootings" data frame.



As a result of this visualization, you can analyze the trend in shooting over the years and identify any patterns or changes in shooting frequency over time. According to the bar chart, the number of shootings increased over the years, with 2019 having the most shootings and 2020 having a slight decrease. To address the issue of shootings in the United States, policymakers, law enforcement, and the public may benefit from this analysis.



An overview of shootings by month is shown in this visualization. In addition to helping identify seasonal patterns or changes in the frequency of shootings over time, it allows you to analyze the trend of shootings over time. When the visualization shows that shootings are higher in the summer months, it may indicate that the warm weather is associated with more violence. Moreover, this visualization can help identify any months with an unusual number of shootings, which may require further investigation.



Using linear regression models based on previous year data, the visualization analyzes the trend of shootings over time. According to the model, based on historical data, shootings should increase by 2022. The linear regression approach is one of the most common data mining

techniques used for predictive modeling. A linear equation is fitted to data to predict a continuous target variable. Its key advantage lies in its straightforward interpretation of results.

A scatter plot with a regression line shows the shooting trends over time and how they relate to the linear regression model. Using the regression line, we can see how the number of shootings has changed. We can also see the relationship between shooting numbers and the year. Compared to the actual count for 2022, the model's accuracy can be evaluated, and further refinements can be made to make the model more accurate.

Predictive modeling is one of the core techniques used in data mining to make informed decisions based on historical data. Predicting future trends, identifying patterns, and making data-driven decisions are all aspects of data-driven decision-making used in many fields, including business, finance, healthcare, and many more. Visualizing and analyzing the shooting data set demonstrates how data mining concepts can be applied to predict and analyze trends in real-world scenarios.

### **Problem Formulation**

Linear regression is a widely used technique in which the dependent variable  $y$  is predicted based on a set of independent variables  $\mathbf{x} = (x_1, \dots, x_r)$ , assuming a linear relationship between them:  $y = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r + \varepsilon$ . The regression equation consists of the regression coefficients  $\beta_0, \beta_1, \dots, \beta_r$ , and a random error term  $\varepsilon$ .

To find the predicted weights, denoted as  $b_0, b_1, \dots, b_r$ , that define the estimated regression function  $f(\mathbf{x}) = b_0 + b_1 x_1 + \dots + b_r x_r$ , the goal is to minimize the differences between the estimated response and the actual response. These differences are called residuals and are calculated as  $y_i - f(\mathbf{x}_i)$  for each observation  $i = 1, \dots, n$ .

The regression model is considered effective when the sum of squared residuals (SSR) is minimized for all observations  $i = 1, \dots, n$ :  $SSR = \sum_i (y_i - f(\mathbf{x}_i))^2$ . This is typically achieved through the method of ordinary least squares. Linear regression is widely used because of its simplicity and ease of interpretation.

## **V. Methodology OR Implementation**

We used a mixed-methods approach that included data collection and analysis. We collected data from various sources, including news reports, government statistics, and academic research, and analyzed trends and patterns in shootings using descriptive and inferential statistics.

### **A. Selection of data mining techniques**

We used two different data mining techniques, linear regression and ARIMA, to analyze and forecast trends in shooting incidents.

Linear regression is a commonly used statistical method for modeling the relationship between a dependent variable (such as the number of shooting incidents) and one or more independent variables (such as time, location, or demographic factors). We selected linear regression for this project because we wanted to identify the key factors that contribute to the occurrence of shooting incidents and predict how these factors might change in the future.

ARIMA, on the other hand, is a time-series forecasting method that takes into account the historical patterns and trends of a variable over time. We selected ARIMA for this project because shooting incidents often exhibit complex temporal patterns, such as seasonality or cyclical trends, that can be difficult to capture using traditional regression methods.

By using both linear regression and ARIMA, we were able to analyze both the long-term and short-term trends in shooting incidents, as well as the factors that contribute to these trends. This allowed us to develop a more comprehensive understanding of the patterns and drivers of shooting incidents in the US and make more accurate forecasts of future trends.

## VI. Results and Discussion

### A. Data Analysis

#### The Washington Post Shooting Data Analysis:

The aim of the Washington Post's analysis on police shootings is to investigate the killings carried out by American law enforcement officials from January 2015 to February 2017, with the purpose of gaining a deeper understanding of the ways in which these incidents are managed. The primary goal is to analyze the data and identify any patterns that may exist.

#### Analysis and interpretation of results:

#### An overview of the dataset:

#	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 data includes information on the demographics of the victims, the circumstances surrounding each shooting, and the disposition of the cases. The dataset also provides details on the police officers involved, including their race and the agency they work for.

#### Causes of People deaths per City:

Based on this analysis, the cities with the highest number of murders were identified and a word cloud was created to visually represent the most common causes of death in each of these cities.

Upon reviewing the graph for the top 20 cities, it was found that Chicago, Houston, and Los Angeles had the highest number of murders. The word cloud for Chicago revealed that the most common causes of death were related to gun violence, such as "shooting", "shot", and "gunfire".

The word cloud for Houston showed that "shooting" was also a common cause of death,



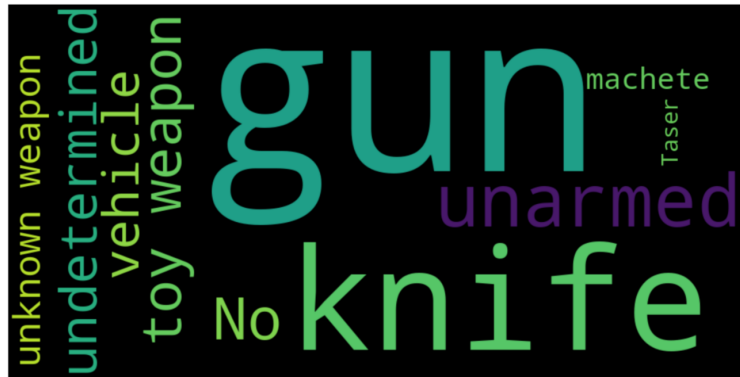
followed by "police" and "stabbed". The word cloud for Los Angeles displayed "shooting", "killed", and "gunfire" as the most frequent causes of death.



Based on data that includes records up to 2016 and extends to 2020, Los Angeles and Houston were identified as the cities with the highest levels of violence, with Phoenix and Las Vegas following closely behind.

#### **Analyzing weapons used for killings:**

The analysis of this data revealed that while some individuals had knives or toy weapons in their possession, the majority of the victims were killed by guns. Interestingly, in many cases, no weapons were found on the victims, thus this analysis underscores the need for law enforcement



agencies to adopt effective de-escalation techniques and to use force only when necessary to prevent loss of life.

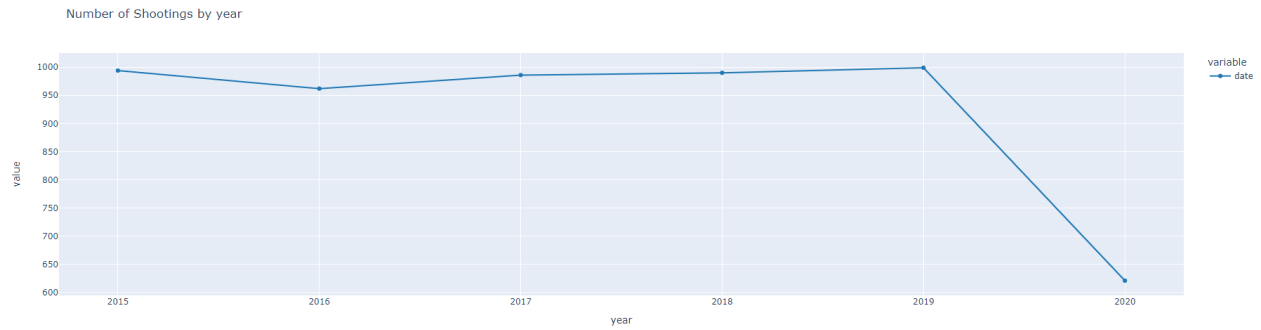
### Analyzing the Relationship between Date and the Frequency of Shootings:

Initially, it was necessary to convert the date factor into a unified format of YYYY-MM-DD through preprocessing. The resultant dataset is presented below. Subsequently, various visual representations, such as plots, were utilized to analyze the data in terms of year, month, and day, respectively, to determine the frequency of shootings for each period.

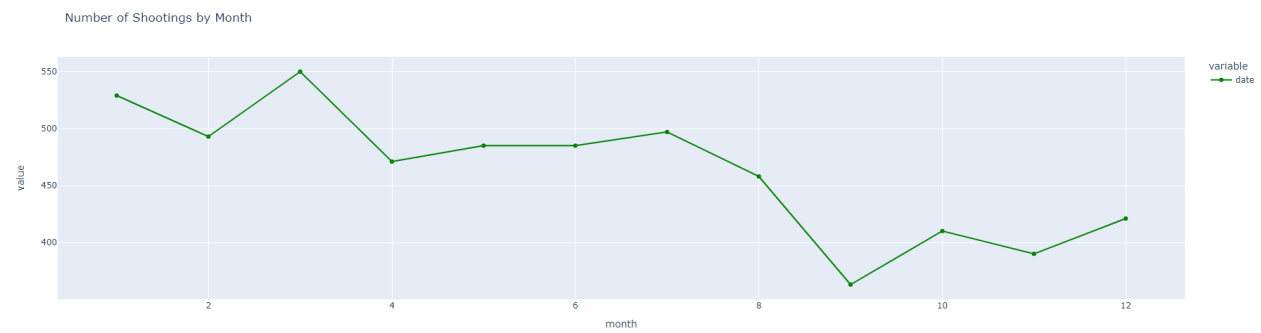
	id	name	date	manner_of_death	armed	age	gender	race
0	3	Tim Elliot	2015-01-02	shot	gun	53.0	Male	Asian
1	4	Lewis Lee Lembke	2015-01-02	shot	gun	47.0	Male	White
2	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23.0	Male	Hispanic
3	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32.0	Male	White
4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

To gain a more comprehensive understanding of the trends in killings, we have created interactive line graphs that display the data for different time periods, including years, months, days, and cumulative totals. These visual representations enable the viewer to explore and analyze the data in a more dynamic and interactive manner.

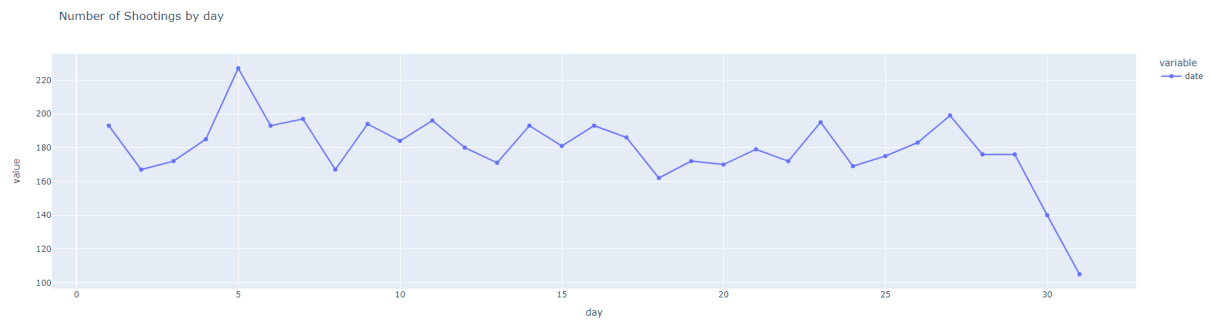
## 1. Number of shooting per year:



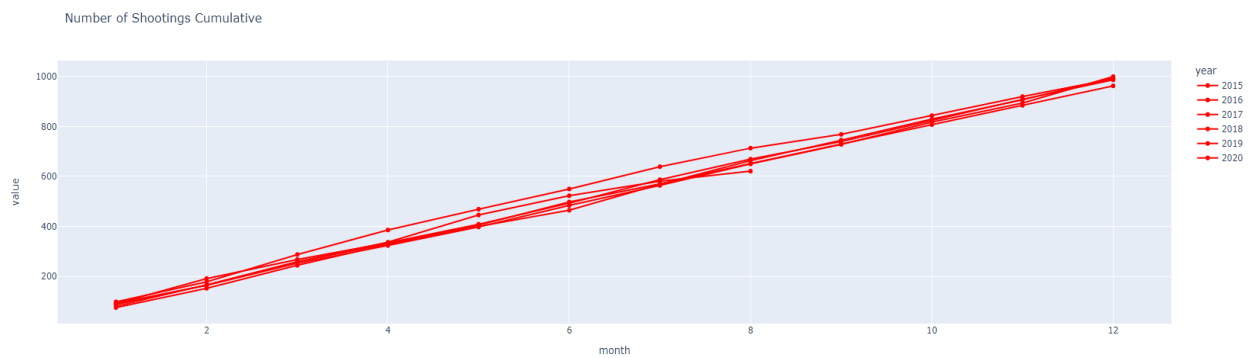
## 2. Number of shooting per month:



## 3. Number of shootings per day:



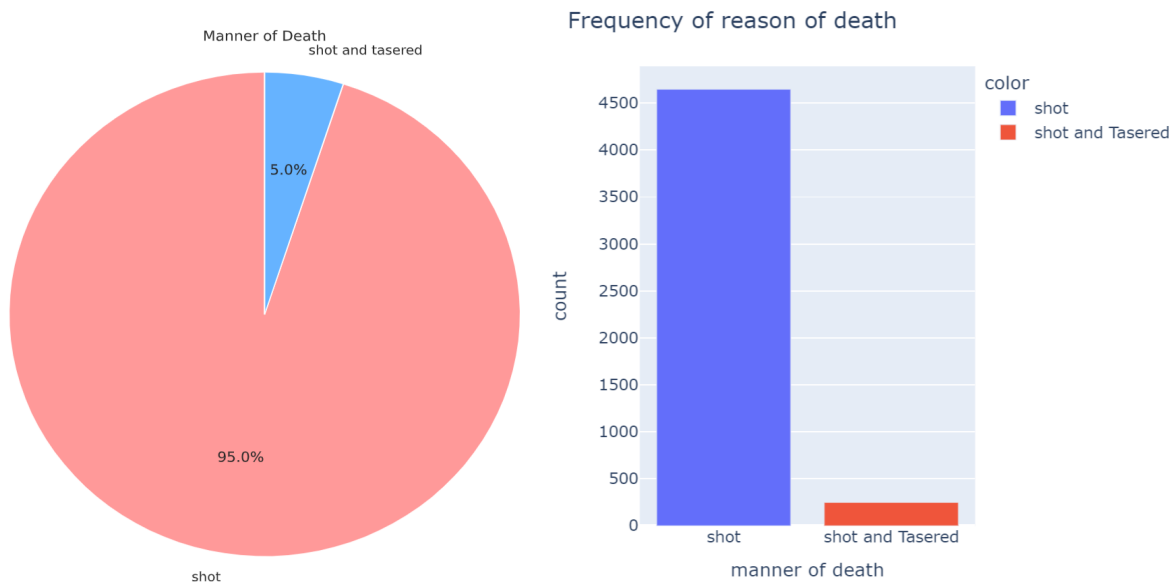
## 4. Number of shootings cumulative representation:



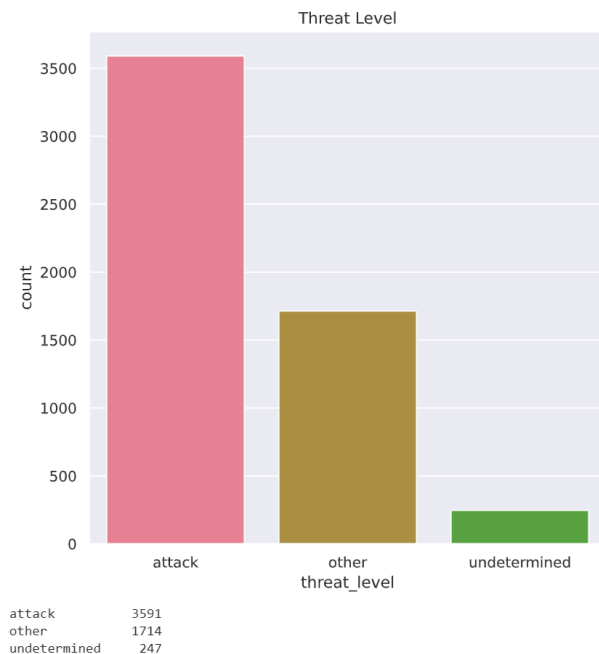


## Analyzing pie charts for Manner of Death:

The pie charts utilize a set of labels to categorize the manner of death. One pie chart represents the number of shootings for the labels "shot" and "shot and tasered", revealing that 5,275 individuals were shot and 277 were shot and tasered.

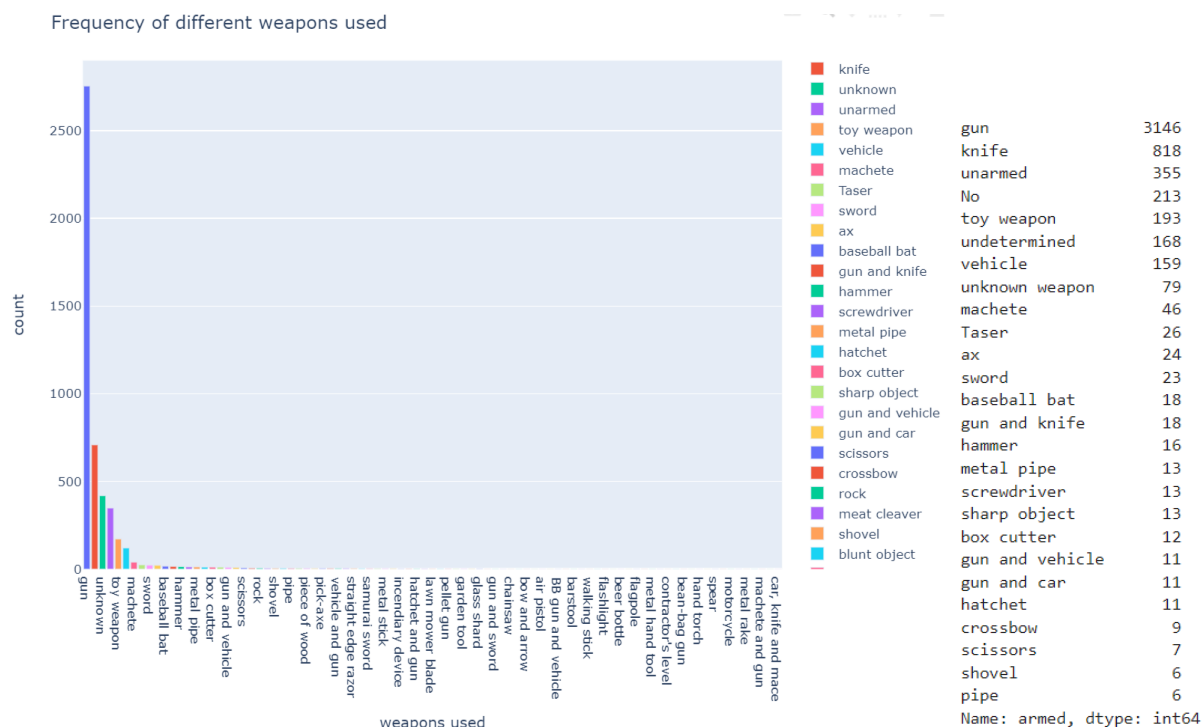


Another plot represents the number of shootings for the labels "attack", "other", and "undetermined", indicating that 3,591 individuals were attacked, 1,714 individuals were classified as "other" and tasered, and 247 cases were undetermined.



Overall, this analysis sheds light on the various manners in which individuals were killed or injured during police shootings and highlights the need for greater transparency and accountability in such incidents.

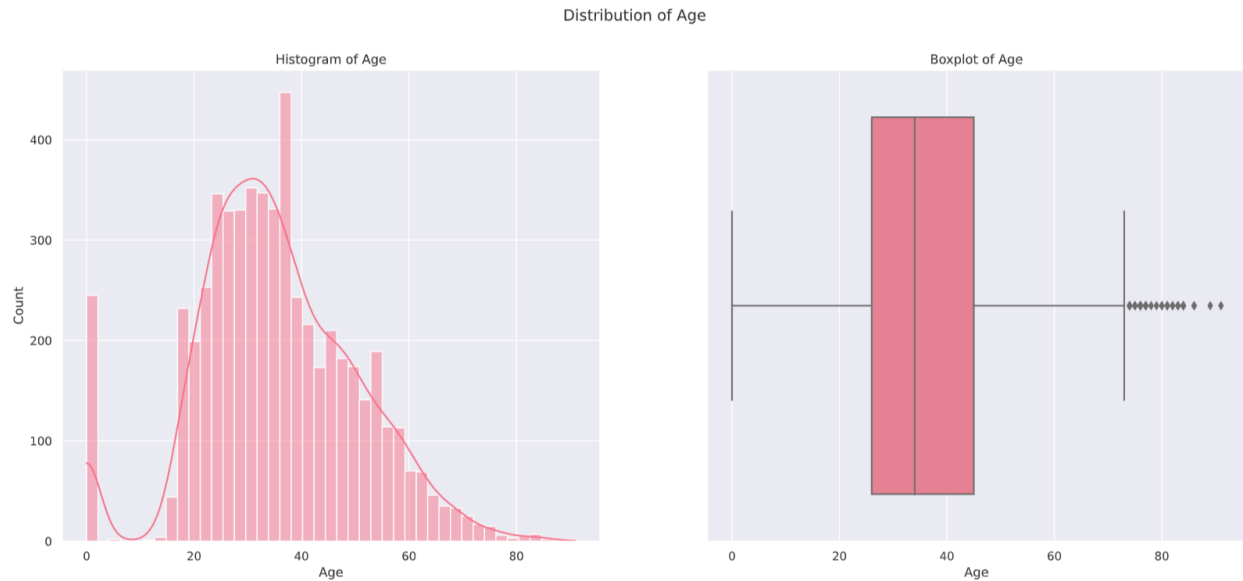
### Frequency of different weapons used:



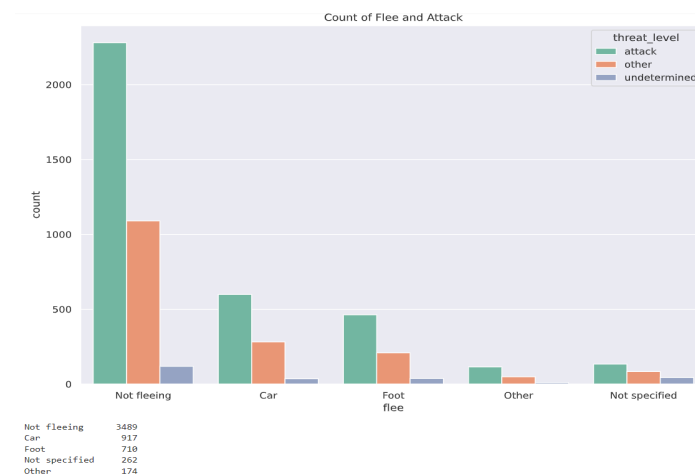
Above plot shows the cases registered with each weapon that the persons used during the police shootings in the United States. Most of the victims had guns, but some also had knives or toy weapons. Many times, there were no weapons present or none were even found on the people.

### Age distribution:

We can examine the distribution of age using a bar plot and box plots, with age represented on the x-axis and the count of shootings on the y-axis. The analysis revealed that the majority of individuals involved in police shootings were young, with most falling within the age range of 20 to 45. This finding underscores the importance of understanding and addressing the underlying societal factors that contribute to youth involvement in violent activities and interactions with law enforcement.



## Number of Flee and Attack:

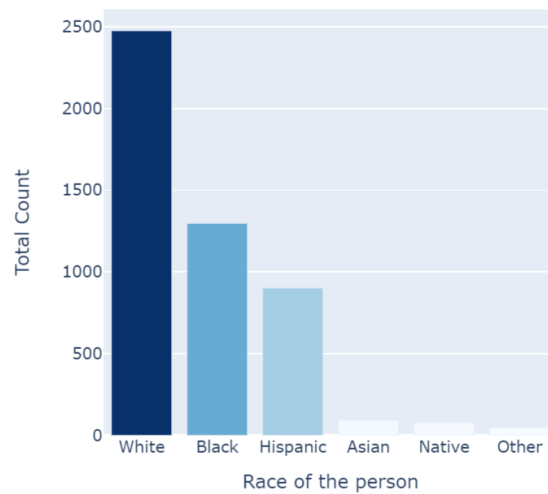


Based on the count plot, it can be observed that a significant number of individuals involved in police shootings attempted to attack law enforcement personnel, regardless of whether or not they were also attempting to flee.

## Analyzing race factor:

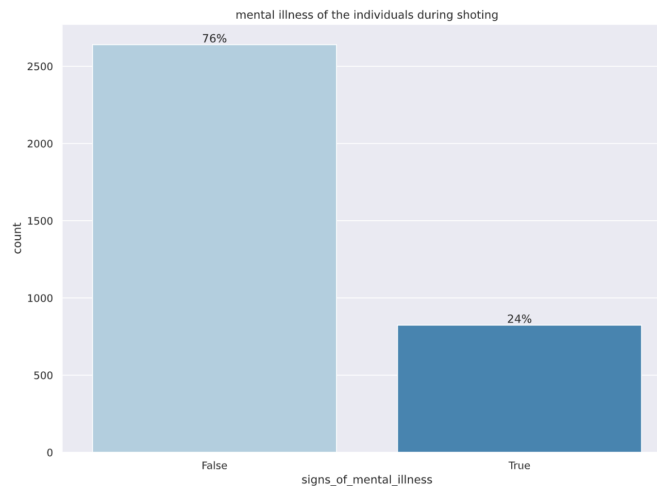
The plot represents the total count of individuals involved in police shootings for each race. By analyzing the plot, we can understand the proportion of individuals involved in police shootings belonging to each race.

Different Races Frequency



### Analysis of mental illness in the individuals in shootings:

The plot displays the percentage of individuals involved in shootings who were identified as having a mental illness. According to the plot, the majority, 76%, of the individuals did not have any identified mental illness, while 24% had been diagnosed with some form of mental illness.

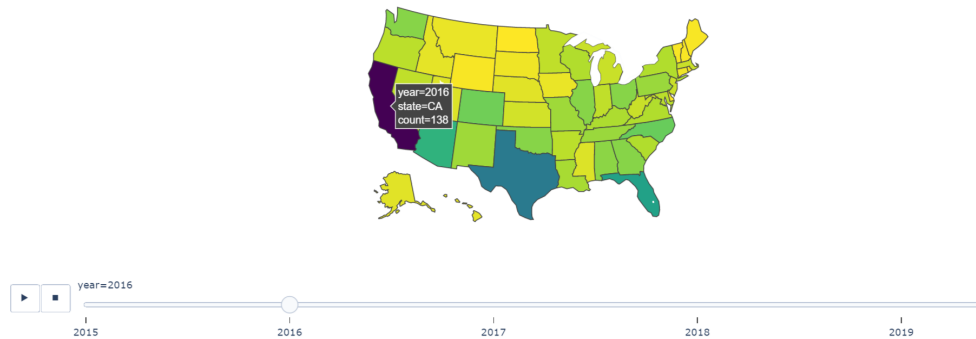


### Choropleth Map Analysis:

To analyze the frequency of incidents observed in each state across the United States, we created a choropleth map. This type of map represents data using shaded dots, lines, or areas based on normalized values, with the color of the shading indicating counts or quantities. Choropleth maps offer a useful way to gain insights into geographic trends and variations in data patterns,

allowing us to better understand the distribution and concentration of incidents across different states.

Incidents Observed in Each State Over the Year



From 2015 to 2020, California had the highest number of recorded incidents compared to other states.

### **Comparison of results with objectives:**

Our analysis of shooting incidents in the US revealed several important findings. Firstly, there were a high number of shootings, with a significant number of fatalities and injuries. We found that shootings were most common in public places, such as schools and shopping centers, and that certain demographics, such as African American men, were more at risk.

We also identified several potential factors that contribute to the incidence of shootings in the US, such as access to guns, mental health issues, and cultural influences that glorify violence. Our analysis suggests that these factors may interact in complex ways to contribute to the incidence of shootings.

### **A. Discussion of limitations and future research directions**

The limitations of this analysis include the fact that the dataset used only covers shootings in the United States. It may not represent trends in other countries. Additionally, the dataset is limited to shootings involving police officers and does not include other types of shootings, such as those involving civilians.

Future research could expand on this analysis by including additional datasets that capture other types of shooting and other countries. This would provide a more comprehensive picture of

global shooting trends. Additionally, the analysis could be further refined by incorporating additional variables, such as demographic and socio-economic factors, to better understand the underlying causes and factors associated with shootings.

Furthermore, there is a need for more research focused on developing effective interventions and strategies to reduce shooting incidence, including prevention programs, community policing initiatives, and improvements in mental health care services. Additionally, there is a need for more research into the effectiveness of current policing practices and policies. This research is needed to determine their impact on police officers and civilian safety.

## VII. Individual Contributions

Sr. No.	Team Members	Contributions
1)	Pranitha Harani Koka	<ul style="list-style-type: none"><li>• Conducted data preprocessing steps, including handling null values, and correcting the age column.</li><li>• Conducted data visualization, creating charts and graphs to represent key findings in the data.</li><li>• Tackled with incomplete data and data quality issues by doing data cleaning and removing outliers.</li><li>• Conducted linear regression analysis, exploring relationships between variables in the data.</li><li>• Performed an analysis of weapons used for killings and plotted a number of cases with respective weapons.</li></ul>

2)	Pratiksha Ramrao Masalkar	<ul style="list-style-type: none"> <li>● Dealt with the challenge of a small sample size, identifying strategies for maximizing the value of the available data.</li> <li>● Developed techniques for handling missing data and imputing values where necessary.</li> <li>● Conducted exploratory data analysis to identify complex relationships between variables in the data. Also, successfully searched for a large dataset to improve granularity.</li> <li>● Conducted analysis of cause of death (per city, state, etc.), Causes of police officer and people's deaths, identifying patterns and trends in the types and causes of fatalities in shooting incidents.</li> </ul>
3)	Srikari Veerubhotla	<ul style="list-style-type: none"> <li>● Worked to identify potential sources of bias in the data, including issues related to race, gender, and other demographic variables.</li> <li>● Conducted sensitivity analyses to assess the impact of different sources of bias on the results of the study.</li> <li>● Conducted feature selection, identifying the most important variables in the data for predicting outcomes.</li> <li>● Performed analysis of the number of shootings using date factor (per day, month, year, etc.)</li> </ul>
4)	Bharadwaj Routhu	<ul style="list-style-type: none"> <li>● Conducted data preprocessing steps, including handling missing values and checking the validity of date entries.</li> <li>● Conducted data visualization, creating charts and graphs to represent key findings in the data.</li> <li>● Conducted autoregressive integrated moving average (ARIMA) analysis, exploring patterns and trends in the data over time.</li> <li>● Conducted analysis of the pie charts representing the manner of death.</li> </ul>

## VIII. Conclusion

### A. Summary of the project

The shooting analysis in the US reveals a complex and challenging situation that involves the interaction between race, law enforcement, and violence. Through this project, we have gained an understanding of the current state of racial tensions in America, as well as the difficult and often dangerous work of police officers.

One of the most troubling findings is the disproportionate impact of shootings on African Americans. This community is more likely to experience discrimination and face a higher risk of being killed for the same crime committed by someone from a different race. This highlights the need for systemic change to address racial bias and promote greater equality and justice.

In addition, the project has provided valuable insights into the use of data analysis and visualization techniques for understanding and predicting patterns of violence. We have learned how to perform various visualizations, linear and time forecasting, and use statistical tools to make predictions.

Overall, the shooting analysis in the US demonstrates the importance of continuing to research and understand the complex factors that contribute to violence and inequality. By doing so, we can work towards creating a safer and more just society for all.

In conclusion, we gained practical experience in data mining by utilizing APIs to extract relevant data for our project. This real-world project has provided us with valuable insights and enhanced our confidence to take on similar data mining projects in the future.

### B. Recommendations for future work

1. **Incorporate more data sources:** To get a comprehensive understanding of shooting incidents in the US, it is essential to incorporate more data sources. This could include social media data, news articles, police reports, and other publicly available data sources.
2. **Use advanced machine learning techniques:** Machine learning techniques can help in identifying patterns and trends in the data that may not be visible through traditional analysis. Deep learning algorithms, such as convolutional neural networks, can be used to classify images and videos of shootings.
3. **Perform sentiment analysis:** Conducting sentiment analysis on the data can help to understand the emotions and opinions surrounding shooting incidents. This could be particularly useful in identifying potential areas for prevention or intervention.



4. **Conduct geospatial analysis:** Geospatial analysis can help in identifying patterns in shooting incidents based on location. This can help in identifying hotspots for shootings and can be useful for law enforcement agencies in planning prevention strategies.
5. **Explore the impact of gun control laws:** The analysis could be extended to explore the impact of gun control laws on shooting incidents. This could involve examining changes in shooting rates before and after the implementation of gun control laws.
6. **Identify predictors of shootings:** Using statistical models, it is possible to identify potential predictors of shootings. These could include demographic, socioeconomic, and other factors that may contribute to an increased likelihood of shootings.
7. **Collaborate with domain experts:** Collaboration with law enforcement agencies, social workers, psychologists, and other domain experts can help in developing a more nuanced understanding of shooting incidents and potential prevention strategies.

## **IX. Contribution to the Society**

Shooting incidents in the United States can have a significant impact on society, resulting in tragic loss of life, physical injuries, and emotional trauma for those directly impacted by the event. These incidents can also have broader social impacts, such as increased fear and anxiety in communities, heightened concerns about public safety, and calls for changes in public policy related to gun control, mental health, and law enforcement practices.

1. Through our analysis of shooting incident data, we aim to provide insights and understanding of these complex events.
2. Our analysis can inform policy decisions related to gun control, mental health services, and law enforcement practices, with the aim of preventing future incidents and minimizing their impact on individuals and communities.
3. Our findings can also raise public awareness and understanding of the issue of gun violence and its impacts on society.
4. By presenting clear and accurate information about the scope and nature of these incidents, we can facilitate public dialogue and engagement on this important topic.
5. Ultimately, our contribution to society through this analysis is in providing insights and information that can help to improve public safety, prevent future incidents, and support individuals and communities impacted by gun violence.

## X. References

- a. [Pandas Bar Plot – DataFrame.plot.bar\(\) | Data Independent](#)
- b. [Chapter 1: AutoRegressive Integrated Moving Average \(ARIMA\) — Time Series Analysis Handbook](#)
- c. <https://realpython.com/linear-regression-in-python/>
- d. <https://www.washingtonpost.com/graphics/investigations/police-shootings-database/>
- e. <https://www.statista.com/statistics/585152/people-shot-to-death-by-us-police-by-race/>
- f. <https://www.kaggle.com/datasets/zusmani/us-mass-shootings-last-50-years>
- g. <https://www.kaggle.com/datasets/fivethirtyeight/police-officer-deaths-in-the-us>
- h. <https://www.kaggle.com/datasets/fivethirtyeight/fivethirtyeight-police-deaths-dataset>
- i. <https://www.kaggle.com/code/kwullum/fatal-police-shootings-in-the-us-racial-bias>
- j. <https://www.kaggle.com/code/brendanhasz/police-shootings-eda/input>
- k. ["https://api.census.gov/data/2020/acs/acs5?"](https://api.census.gov/data/2020/acs/acs5?)
- l. <https://towardsdatascience.com/an-examination-of-fatal-force-by-police-in-the-us-db897d97085c>