From Data to Insights:
Examining Fatal Police Shootings in the U.S.

Co-Authors:

Sakthi Swarup Vasanadu Kasi- 02085794

Shalabh Singh Yadav – 02130778

Teja Naga- 02122966

The Issues:

In the United States, fatal police shootings are a serious concern, affecting public safety. This challenging topic requires an in-depth analysis that goes below the surface to examine the people, situations, and root causes involved in such instances. Our project aims to gain a complete understanding of these events to guide the policies and actions aimed at improving community safety and police-community relations. By analysing the various aspects of these shootings, we aim that we will develop solutions that promote a safer, more trusting society in which the police work in collaboration with the citizens it serves.

We address the following questions:

1. What is the age distribution of individuals fatally shot by police?
2. How do fatal police shootings vary by race and gender?
3. Is there a correlation between the geographic location (latitude and longitude) and the frequency or nature of these incidents?
4. Are there any notable trends or patterns in the data when analysed statistically, such as clustering of incidents in certain areas or among certain demographic groups?
5. What are the most common scenarios in police shooting incidents, and how do they vary in different clusters identified through clustering analysis?
6. Could predict whether an incident involved a certain type of weapon, whether the suspect had a mental illness, or whether a body camera was present.
7. Predict the nature of the threat posed by the suspect (e.g., aggressive, fleeing) based on variables like the weapon type, mental illness, and demographics.
8. Based on clustering results, can we identify profiles that are at a higher risk of being involved in police shootings?

Findings:

The in-depth analysis of fatal police shootings in the United States reveals unique variations and patterns on the bases of race, gender, and geographic location. The data indicates that a higher incidence happen among white individuals, followed by black and hispanic populations, with a significant gender disparity showing more men were involved in these incidents. This difference is influenced by cultural and behavioural factors. Geographically, a well-focused contrast exists between urban areas like Los Angeles and Houston, which report higher rates of police shootings, and scattered, rural regions with fewer incidents. This inequality suggests hidden social and economic challenges in denser areas. Through K-Mean clustering analysis, the study further straightens outs unique spatial patterns in states like California, Texas, and Florida, indicating that these incidents are not evenly distributed and tend to concentrate in specific regions, both urban and rural.

In predictive analysis, we utilized logistic regression to predict various aspects of the incidents. The model was moderately successful in predicting gun-related incidents, but less so for non-gun-related cases. It also showed bias in predicting the absence of mental illness, reflecting a challenge in recognizing cases with mental health issues. Similarly, it accurately predicted scenarios without body cameras but struggled to identify instances where they were present. Notably, the model's inability to accurately predict aggressive threats, despite its high accuracy in recognizing non-aggressive threats, points to the complexity of these incidents and the

limitations of predictive modelling in fully capturing the fine distinction of fatal police shootings. This interconnected data highlights the multifaceted nature of police shootings, shaped by a combination of demographic, geographic, and situational factors.

Discussion:

The analysis of fatal police shootings in the United States highlights significant and complex implications. It reveals a clear racial and gender disparities, mainly affecting young males from white and black communities. This calls for an in-depth examination of Communicative and systemic factors contributing to these inequalities. The higher incidence of these shootings in urban areas, particularly in states like California, Texas, and Florida, points towards the influence of law enforcement practices in densely populated regions.

To begin with, the clustering of incidents in specific areas suggests that community relations and local governance have a significant impact on the occurrence of these shootings. This Geographical concentration opens possibilities for targeted interference and community-based strategies to mitigate such incidents.

Primarily, the limitations of predictive models like k-means clustering and logistic regression in accurately of these events highlights their unpredictability and complexity. This highlights the need for more complex and comprehensive data collection and analysis methods to understand and anticipate the factors leading to fatal police shootings.

In addition, the analysis also finds that scenarios involving an accused individual, who is armed and does not attempt to flee, constitute a significant proportion of cases. These situations tend to escalate rapidly, increasing the likelihood of a shooting. Additionally, the act of fleeing, even when the individuals involved are unnamed, is a factor that can precipitate a shooting.

A significant limitation encountered in the analysis is the insufficient data, which Impairsthe accuracy of predictions made using logistic regression. This lacksin highlights the need for more comprehensive and higher-quality data to gain a deeper understanding of these incidents.

The study highlights the importance of reforming policies, increasing community involvement, improving police training in de-escalation and racial awareness, and strengthening mental health support to address police shootings. These actions are vital for tackling the root causes of such incidents and ensuring safer, more equitable communities. The findings serve as a foundation for informed discussions and effective strategies to address this critical issue.

Appendix A: Method

The primary intention of the project is to acquire a thorough understanding of all fatal police shootings that have taken place in the US since January 1, 2015. To do this, we access information from the Washington Post, which has a large archive of these types of incidents. The process of encounters has several phases, all of which are essential for deriving significant insights from the data.

1. Data Collection and Initial Setup

   The project started by acquiring the dataset which included detailed statistics on fatal police shootings in the US beginning on January 1, 2015 from the Washington Post. This dataset contains specific dates and locations of each occurrence, along with additional relevant details, as well as detailed information about the individuals involved. Using the panda's library, the data was systematically imported into a Python environment. This was an essential step in facilitating efficient data processing and further analysis.

2. Data Cleaning and Pre-processing

   The dataset underwent a thorough cleansing procedure. Several instances of missing numbers, inconsistencies, and errors were found after a thorough investigation. To maintain the general integrity of the analysis, important judgments had to be made regarding the deletion of specific items where the data was missing or considered unreliable. Pandas was a tool we used a lot to find incomplete data or missing data. With consideration of important variables like "state," "age," "race," "gender," and geographic locations, this procedure was particularly stringent. The data underwent a series of transformation and normalization processes, standardizing various categories and ensuring a level playing field for comparison and analysis.

3. Exploratory Data Analysis (EDA)

   The EDA stage included an in-depth collection of quantitative and qualitative analyses. A wide range of statistical techniques were used to analyse the data and understand the patterns, with a particular emphasis on demographic details and the temporal spread of the incidents. The collection of bar plots and other graphical representations were created to provide a visual explanation of data distributions such as incidence frequencies classified by state, race, and other demographic factors. These visual tools were essential for finding root causes and defects offering insights that would probably have gone undiscovered in a purely numerical investigation. We created graphs and plots using computer programs (matplotlib and seaborn). We employed encoding techniques to handle categorical data. This involved transforming non-numeric categories into numeric values so that they could be used.

4. Geographical Analysis and Visualization

   One important element of this investigation was geographic analysis. Detailed maps were created using the spatial data, and advanced visualizations have been generated with the help of tools like Folium. These maps highlighted global patterns, variations among regions, and potential hotspots of incidents, making them informative as well as illustrative.

The spatial aspect of the research offered an entirely new viewpoint on the data, enabling the identification of regional trends and connections.

5. Advanced Statistical Analysis and Machine Learning Techniques

   A section of the analysis was allocated to more advanced analytical approaches. To validate the models against unseen data, the dataset was divided into training and testing sets, as is standard procedure in machine learning.

   - Training and Testing Split: We used the train_test_split function from scikit-learn to divide the dataset into training and testing subsets. This allowed us to train our models on one part of the data and test their performance on a separate set, ensuring the durability of our models.

   Key methods applied included:

   - K-Means Clustering: To identify different patterns or groupings in the incidents, we performed K-Means clustering to identify naturally occurring groupings within the data. This method made it easier to determine whether incident types were more likely to happen clusters—that is, in particular locations or situations. A crucial step in this approach was figuring out how many clusters would be optimal as this helped us classify the incidents and identify underlying trends that a simple analysis might not have picked up immediately.
   - Logistic Regression: This method was used to investigate the impact of various factors on the outcomes of police encounters, such as demographics or location. The intention was to determine which factors in these instances where most significantly connected with the chance of a tragic outcome. Logistic regression was useful in estimating the risk associated with the different factors, providing a clearer understanding of the essential components in these incidents.

The project concluded with a summary of the key discoveries and their broader implications. These findings led to the developing a series of suggestions that included potential changes to policy, directions for future research, and techniques for improving police work and community relations.

Appendix B: Results

This project takes a detailed look at cases where police have fatally shot people in the United States. It emphasizes on multiple features, which include the locations of these instances as well as the people involved in it, including the age, race, and gender. This wide collection of instances assists in comprehending the shared traits and unique characteristics of these incidents in diverse states and societies. The age distribution of people who received gunshots by police as well as died is shown in the **Figure-1**. It illustrates the range and frequency of age group [20-40 years] were most affected by these occurrences.

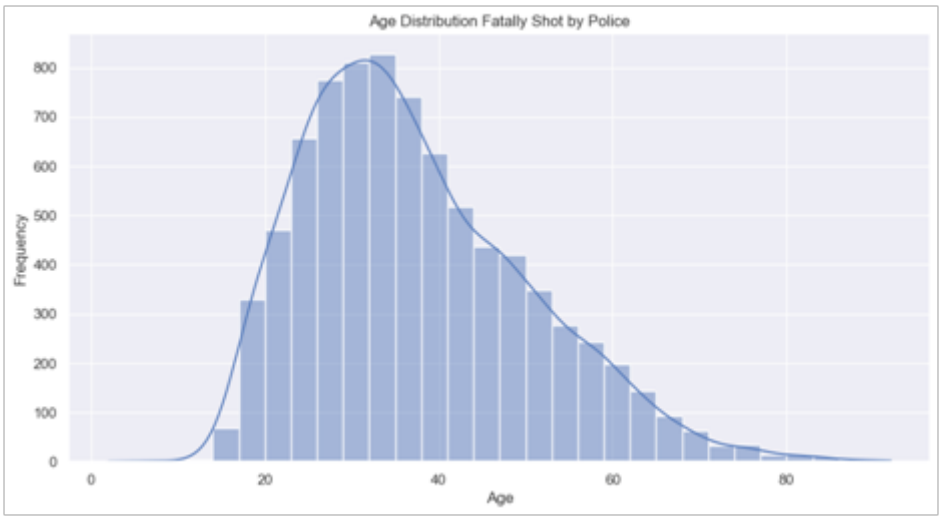


Figure 1: - Age Distribution fatally shot by Police.

When analysing fatal police shootings by gender and race, significant variations were observed. The amount of fatal police shootings illustrates notable racial differences among various racial groups. Data on fatal police shootings frequently show a high proportion of white people. While the data also show a more significant presence of black people when compared to Hispanic people. Though they are fewer numerous, Asian, Native American, and other racial groups are represented. Regarding the results of the gender analysis, the data shows an obvious disparity in the percentage of men and women participating in these instances. There is a significant gender disparity in the frequency of fatal police shootings, which may indicate that cultural, behavioural at play in these incidents. These trends are depicted in **Figure-2**, highlighting the disparities across different demographic groups.

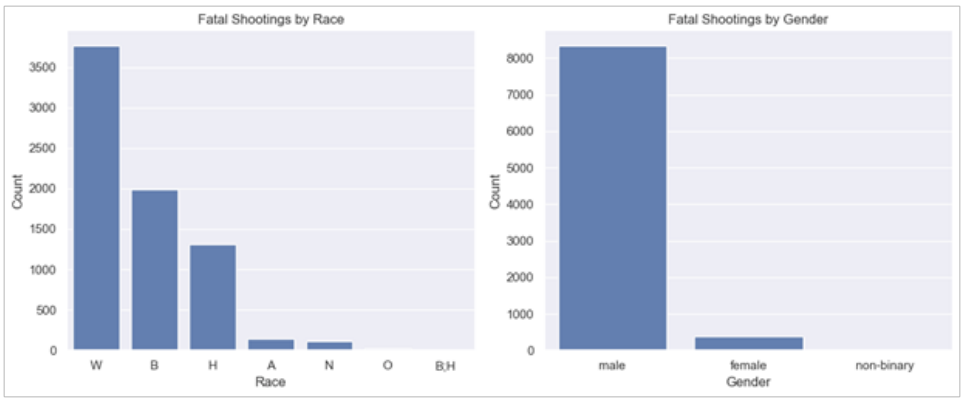


Figure 2: - Variation of Fatal Police Shootings by Race and Gender.

Following an analysis of the differences in fatal police shootings by gender and race, it became essential to analyse these instances on a geographic basis. The geographic distribution and density of these shootings can be shown visually on a map according to latitude and longitude given in the dataset. This analysis provided greater insights into how location connects with gender and race in the context of these incidents, as well as to find any regional patterns or discrepancies as shown in **Figure-3**

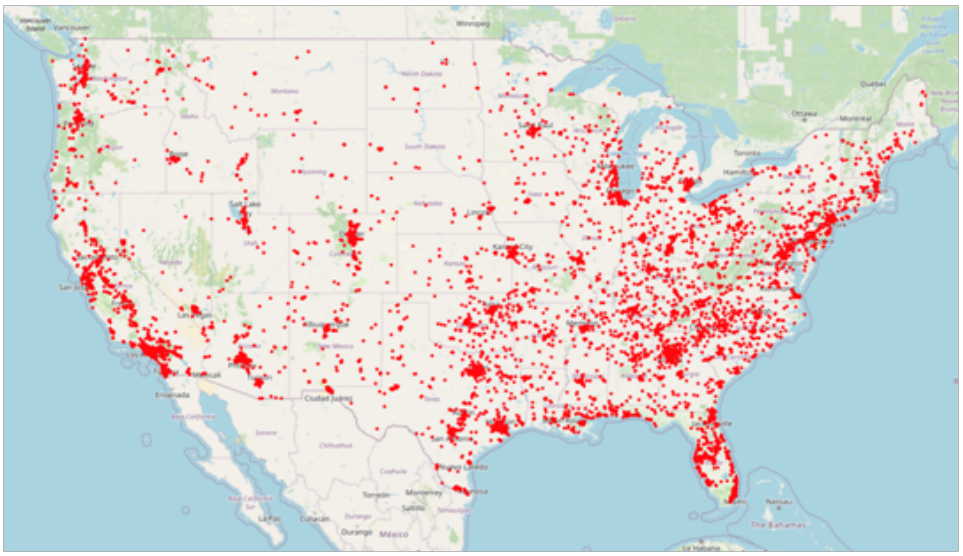


Figure 3: - "Geographical Distribution of Fatal Police Shootings in the United States.

After carefully analysing the map, I found that the distribution of fatal police shootings across the United States exhibits distinct geographic and demographic patterns, which are as follows:

- Urban concentration: - Large, urban cities like Los Angeles and Houston experience greater rates of police shootings. This might be due to a rise in the number of individuals, facing variety of social and economic problems. Police shootings occur more often in the eastern region of the United States, especially in densely populated cities like New York and Pennsylvania, because of increase in population and development in these areas.
- Countryside and Middle of the U.S: -Conversely, there are fewer police shootings in rural areas, the middle of the country, and areas with a lot of mountains and woods. This might be because there are fewer people, compared to big cities, and the police may be required for numerous reasons.

We were able to figure out the patterns and characteristics of these incidents after closely examining the locations. As a result, we decided to focus on California (CA), Texas (TX), and Florida (FL), the three largest states in the United States where cases are more frequent. We employed an approach referred to as K-Mean clustering to assist us classify the incidents into four different spatial groups for each of these states. This method enables us to identify regions where these unfortunate incidents have become more common.

Every state has a unique pattern of its own. Some clusters are spread throughout rural areas, whereas others are concentrated in urban areas. The scatter plots indicate that incidents aren't evenly distributed, instead they focus on specific regions as shown in Figure-4. Bar charts as shown in Figure-5, are used for displaying the total number of incidences in each cluster for every state.

1. California (CA): -

- With 362 incidents, cluster 0 is the largest cluster. It mostly depicts the state's crowded urban areas.
- Comprising 145 events, this cluster 1 illustrates a combination of suburban as well as urban regions.
- With 89 incidences, this cluster 2 is one of the smaller ones, suggesting less occurrences in these areas.

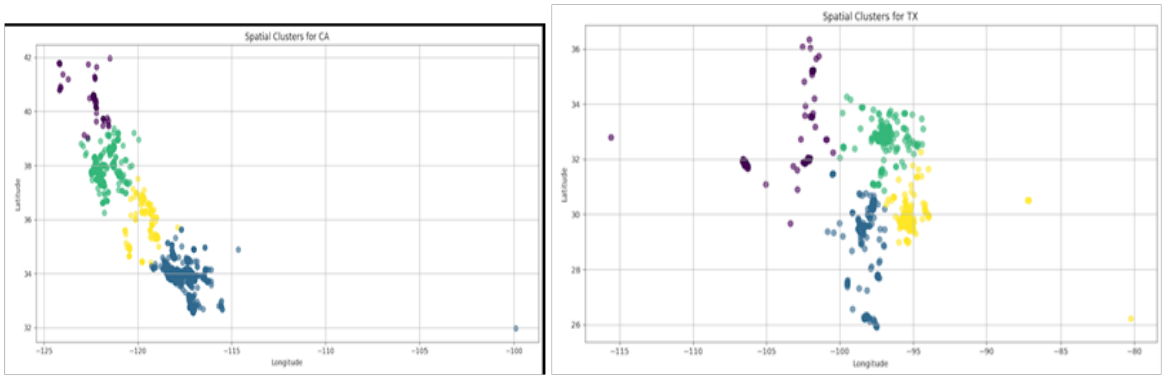
- Cluster 3, which includes 239 occurrences, is spread across multiple urban zones

2. Texas (TX):

- With 266 incidents, Cluster 0 is the largest cluster and includes major cities and the areas around them.
- The 115 occurrences in this cluster 1 originate mainly from the state's eastern region.
- This cluster 2, which has 179 incidents, is distributed throughout the central areas.
- With 67 events, Cluster 3 is the smallest cluster in Texas.

3. Florida (FL):

- Cluster 0: This cluster, which includes the northern region, represents 104 incidents.
- Cluster 1: Comprising 249 events, this cluster is the largest in Florida and comprises of southern region, including the Miami area.
- Cluster 2: With just 17 events, this cluster is the smallest.
- Cluster 3: It includes the state's centre regions with 89 incidents.



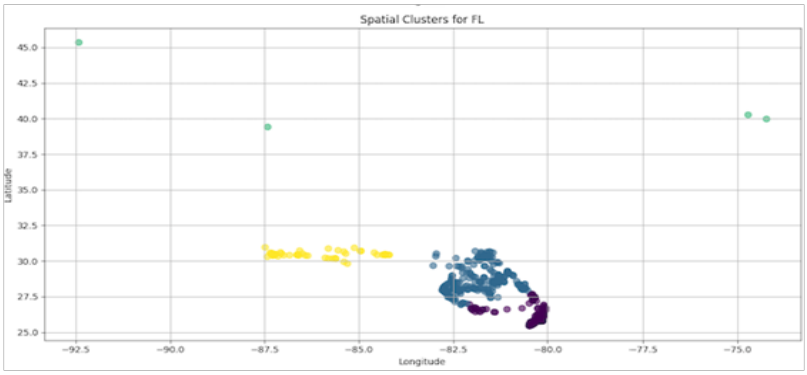


Figure 4: - "Spatial Clusters of CA, TX, FL.

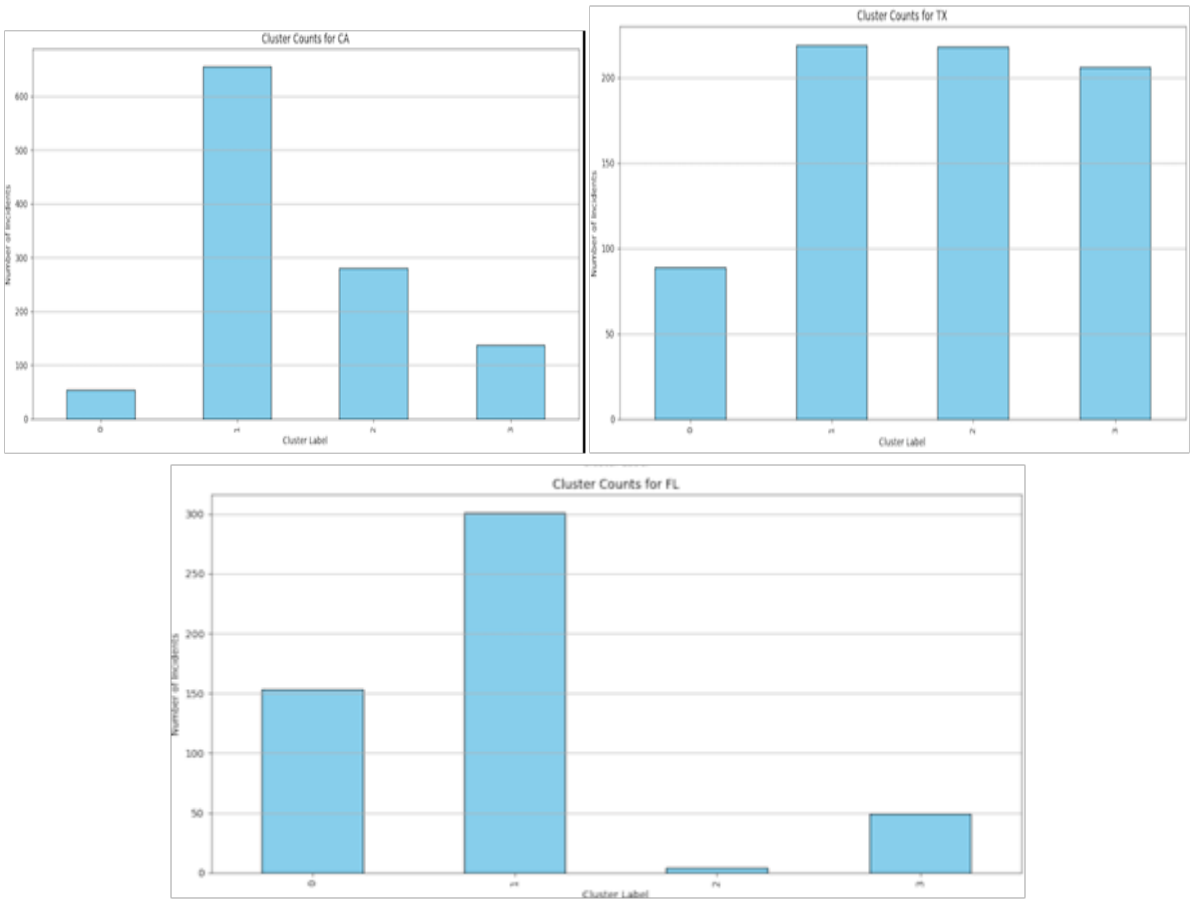


Figure 5: - Cluster Counts for CA, TX, FL.

Understanding the regional patterns and trends of such incident through k-means clustering, it was crucial to understand the most typical police shooting event scenarios which may be at higher risk. So, I performed k means clustering on the dataset. The visualizations in Figure 6 provides a clear picture of the circumstances.

- Cluster 0 comprises events mainly of unarmed individuals who attacked and then fled in a car. with most cases, there was no mental disease involved with these instances.

- Events in Cluster 1 are defined by people who were carrying firearms, did not run away, and behaved in a shooting manner. Additionally, most of these incidents did not include mental illness.

- Cluster 2 is made up of events in which the attackers did not run away, had knives, and had no connection to mental illness.

- Cluster 3 comprises situations in which there was no connection between mental illness and the persons who constituted a danger, stayed put, and carried replica weapons.

- In terms of the threat and the inability to run, Cluster 4 is comparable to Cluster 3, but people were carrying knives, and there was no mental illness.

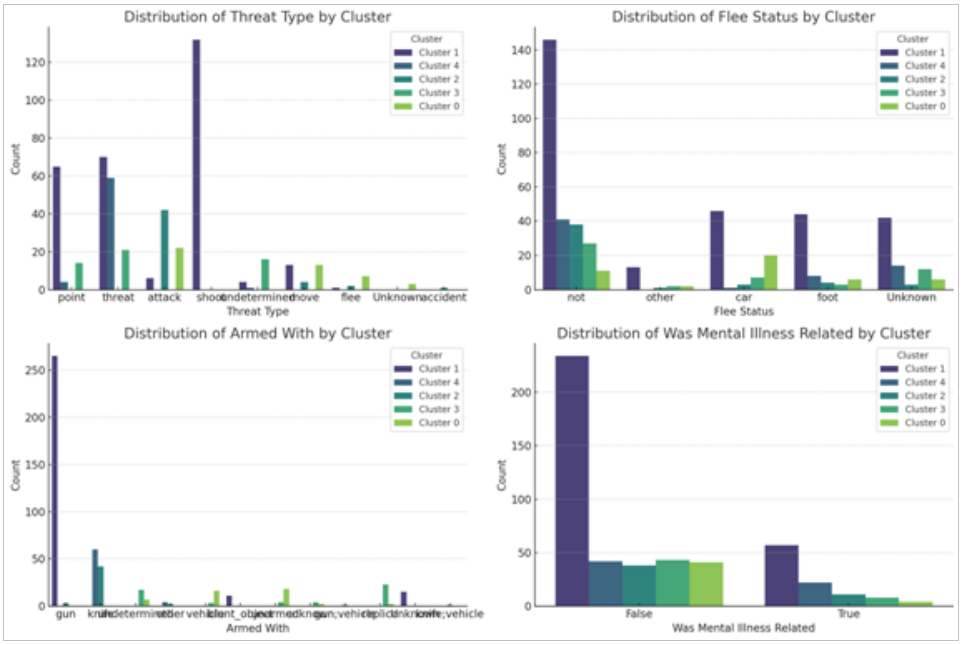


Figure 6–Higher risk to the police

It was crucial to understand whether a certain type of weapon was used, if the suspect had a mental illness, or if a body camera and to predict the threat type during the incident. This was key to predict how these events might occur. So, we used logistic regression for my prediction. The summary of the model can be seen from Figure-7. The results we analysed from the model are as follows:

- Weapon Type Prediction (Gun vs. Non-Gun): The model proved better at predicting gun-related occurrences than non-gun-related ones, with a modest accuracy rate of 60%. Recall for gun-related occurrences was higher (91%) than for non-gun-related incidents (18%).

- Mental Illness Involvement Prediction: Although this algorithm was strongly biased towards predicting the absence of mental illness, it still presented great accuracy (80%). It found it challenging to recognize cases with mental disease (only 1% recollection), whereas it was typically successful in detecting those without mental illness (100% recall).

- Body Camera Presence Prediction: The forecasting ability of the model was considerably limited for this specific outcome, as shown by its high accuracy (84%) in predicting the absence of a body camera but complete failure to detect any instances where a body camera was present.

- Predicting Threat Type (Aggressive vs. Non-Aggressive): With a recall and precision of 0% for the aggressive threat category, the model showed that it was unable to accurately predict any aggressive threats, despite its high accuracy (81%) in recognizing non-aggressive threats.

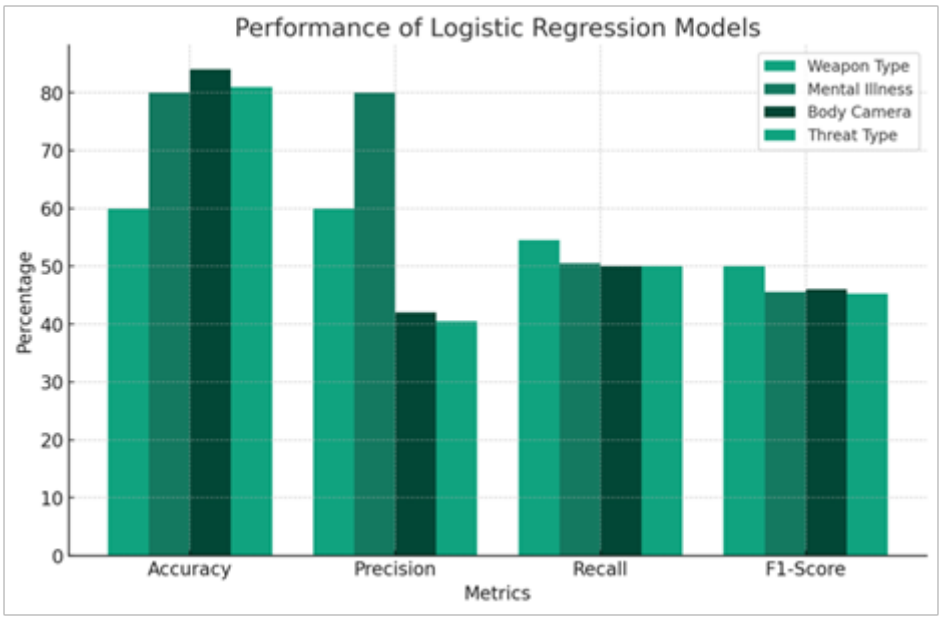


Figure 7 - Performance of the model

The Logistic regression model, in general, showed varying levels of effectiveness, with some constraints in each case. Due to the missing values in the dataset, the model couldn't predict or identify fewer common outcomes (such as the presence of aggressive threats, mental illness, and body cameras).

Appendix C: Code

```
import folium

# Filter out entries with missing latitude or longitude
complete_geo_data = shootings_data.dropna(subset=['latitude', 'longitude'])

# Create a base map
m = folium.Map(location=[37.0902, -95.7129], zoom_start=4)

# Add a circle marker for each shooting
for idx, row in complete_geo_data.iterrows():
    folium.CircleMarker(
        location=[row['latitude'], row['longitude']],
        radius=1, # Small fixed radius for individual events
        color='red',
        fill=True,
        fill_color='red'
    ).add_to(m)

# Save the map to an HTML file
map_file_path = 'map.html'
m.save(map_file_path)

map_file_path
```

```
'map.html'
```

```
# Performing K-Means clustering separately for each state and visualizing the results
fig, axs = plt.subplots(3, 1, figsize=(12, 18))
cluster_counts_states = {}

for i, state in enumerate(states):
    state_data = filtered_data[filtered_data['state'] == state]
    state_coordinates = state_data[['latitude', 'longitude']].dropna()

    # Applying K-Means clustering for each state
    kmeans_state = KMeans(n_clusters=4, random_state=0).fit(state_coordinates)
    state_labels = kmeans_state.labels_

    axs[i].scatter(state_coordinates['longitude'], state_coordinates['latitude'], c=state_labels, cmap='viridis', s=50, alpha=0.6)
    axs[i].set_title(f'Spatial Clusters for {state}')
    axs[i].set_xlabel('Longitude')
    axs[i].set_ylabel('Latitude')
    axs[i].grid(True)

    # Storing the cluster counts for each state
    cluster_counts_states[state] = pd.Series(state_labels).value_counts()

plt.tight_layout()
plt.show()

cluster_counts_states
```

```
# Performing K-Means clustering separately for each state and visualizing the results
fig, axs = plt.subplots(3, 1, figsize=(12, 18))
cluster_counts_states = {}

for i, state in enumerate(states):
    state_data = filtered_data[filtered_data['state'] == state]
    state_coordinates = state_data[['latitude', 'longitude']].dropna()

    # Applying K-Means clustering for each state
    kmeans_state = KMeans(n_clusters=4, random_state=0).fit(state_coordinates)
    state_labels = kmeans_state.labels_

    axs[i].scatter(state_coordinates['longitude'], state_coordinates['latitude'], c=state_labels, cmap='viridis', s=50, alpha=0.6)
    axs[i].set_title(f'Spatial Clusters for {state}')
    axs[i].set_xlabel('Longitude')
    axs[i].set_ylabel('Latitude')
    axs[i].grid(True)

    # Storing the cluster counts for each state
    cluster_counts_states[state] = pd.Series(state_labels).value_counts()

plt.tight_layout()
plt.show()

cluster_counts_states
```

```
# Reversing the label encoding to interpret the clusters
for column, encoder in label_encoders.items():
    circumstances_data[column] = encoder.inverse_transform(circumstances_data[column])

# Grouping the data by cluster and examining the common characteristics
cluster_characteristics = circumstances_data.groupby('cluster').agg(lambda x: x.value_counts().index[0])
cluster_characteristics
```

cluster	threat_type	flee_status	armed_with	was_mental_illness_related
0	shoot	not	gun	False
1	threat	not	replica	False
2	threat	not	knife	False
3	attack	car	unarmed	False
4	attack	not	knife	False

```
# Setting up the aesthetics for plots
sns.set(style="darkgrid")

# Statistical Analysis

# 1. Age Distribution of Individuals Fatally Shot by Police
plt.figure(figsize=(12, 6))
sns.histplot(data['age'].dropna(), kde=True, bins=30)
plt.title('Age Distribution Fatally Shot by Police')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

# 2. Variation of Fatal Police Shootings by Race and Gender
race_counts = data['race'].value_counts()
gender_counts = data['gender'].value_counts()

plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
sns.barplot(x=race_counts.index, y=race_counts.values)
plt.title('Fatal Shootings by Race')
plt.xlabel('Race')
plt.ylabel('Count')

plt.subplot(1, 2, 2)
sns.barplot(x=gender_counts.index, y=gender_counts.values)
plt.title('Fatal Shootings by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')

plt.tight_layout()
plt.show()

# 3. Correlation between Geographic Location and Incident Frequency/Nature
# Calculating correlation coefficients for latitude, longitude, and frequency of incidents
correlation_matrix = data[['latitude', 'longitude']].corr()

correlation_matrix
```

Contribution: -

All four group members analysed the fatal police shooting data. The following task were performed by all the members of the group.

Name	Data Preprocessing	EDA and Stats	Logistic Regression	Kmeans Clustering	Report Writing
Sakthi Swarup Vasanadu Kasi	15	20	40	25	25
Siddharth Jain	15	25	30	35	25
Shalabh Singh Yadav	40	20	15	20	25
Teja Naga	30	35	15	20	25

GitHub link of the project: https://github.com/siddharth00914/Fatal_police_shootings