Behind the Badge: Unraveling the Complex Threads of Police Shootings in America

Akhil Kumar Reddy Bhoomi Reddy - 02080263
Shivakumar Pasem - 02077631
Ben George Samuel - 02078919
Aishwarya Ganesh Bhat - 02097374

1. **The issues**

In navigating the complex landscape of fatal police shootings in the United States, critical questions emerge, demanding thoughtful examination. The Washington Post's meticulous database meticulously chronicles incidents from 2015 onwards, offering a lens into the unsettling dynamics surrounding these events. This comprehensive repository, encompassing details such as age, race, mental illness among the few, lays the groundwork for a nuanced exploration of the issue at hand.

Within this context, our investigation aims to unravel specific dimensions:

- Age Disparities: We scrutinize the data to discern any disparities in the average age of individuals fatally shot by police, differentiating between black and white victims.
- Racial Dynamics: An essential facet of our inquiry delves into the comparison of average ages between black and white individuals subjected to fatal police shootings. This leads us to a crucial question: Are these observed differences statistically significant?
- Mental illness Factor: Our focus extends to understanding whether there is a statistically significant difference in the average ages of individuals with mental illness compared to those without.

By addressing these inquiries head-on, our goal is to contribute to a deeper understanding of the multifaceted challenges surrounding fatal police shootings, paving the way for informed dialogue and potential avenues for reform. The data serves as a compass, guiding us through the complexities of an issue that demands thoughtful consideration and proactive solutions.

2. **Findings**

Upon meticulous observation of the Washington Post Repository data on fatal police shootings in the United States, our analysis reveals significant patterns and trends within the dataset. The age distribution displays a subtle rightward skew, indicating a prevalence of younger individuals involved in these incidents, with a peak observed in the late 20s to early 30s. Notably, males overwhelmingly dominate the dataset, constituting the majority of those involved in fatal police shootings. The dataset also highlights a substantial proportion of armed individuals, with firearms being the primary weapon, followed by incidents involving knives and unarmed individuals.

Delving into the racial dynamics, our examination exposes distinct patterns among different racial groups. White individuals emerge as the largest group in the dataset, followed by Black and Hispanic individuals. A detailed analysis of armament by race indicates consistent trends, with Whites, Blacks, and Hispanics predominantly wielding firearms. A concerning trend emerges among unarmed Black individuals, indicating a notably higher incidence compared to other racial categories.

Temporal patterns in fatal police shootings show an even distribution across days of the week, with a slight increase on Wednesdays and Fridays. In contrast, Sundays tend to exhibit a slightly lower frequency of occurrences. When exploring the intersection of mental health and fatal police shootings, individuals displaying signs of mental illness tend to have a slightly higher median age compared to those without such signs, adding complexity to our understanding of these encounters.

Examining the proportional disparities among racial groups in police shootings unveils stark contrasts. Black individuals are shot at a rate approximately 1.70 times their representation in the U.S. population, while Native Americans and Hispanics face rates of 1.31 and 0.77 times, respectively. Conversely, White individuals experience a rate of approximately 0.67 times, Asians at 0.29 times, and individuals from other racial categories at 0.09 times their respective U.S. population representation.

Perhaps most striking is the finding that the average age of Black individuals fatally shot by police is slightly over seven years less than their White counterparts. This humanizes the statistical analysis, underscoring the need for a nuanced understanding of the factors contributing to these incidents.

3. Discussion

Through comprehensive research and the application of a predictive model, our analysis reveals a noteworthy stability in the number of fatal police shootings, aligning closely with recent observations. This investigation also enabled the identification of major data clusters, unveiling distinct patterns. Notably, Cluster 0, characterized by armed males around the age of 35, and Cluster 2, involving males in their 40s, predominantly feature white individuals. In contrast, Cluster 1, defined by armed males around 33 years old, predominantly involves black individuals, and Cluster 3, consisting of armed females with an average age of 37, predominantly involves white females. Cluster 0 incidents frequently have body cameras present, unlike the other clusters where body cameras are predominantly absent. This nuanced analysis provides a structured perspective on incidents based on selected features and can be refined or expanded based on specific investigative questions or required insights.

Cluster 0 incidents exhibit dispersion nationwide but are notably concentrated in the eastern half of the USA, particularly in the Southeast and Midwest regions. In contrast, Cluster 1 incidents display a more even distribution across the country, lacking concentration in specific regions. Cluster 2 incidents span the entirety of the USA, with pronounced concentrations in densely populated areas and along major highways.

Conversely, Cluster 3 incidents appear more localized, with concentrations observed in regions such as the West Coast and parts of the Midwest.Across all clusters, incidents often involve a threat level categorized as "attack," with individuals typically not fleeing. This insight suggests potential common circumstances leading to most shootings. To refine insights further, deeper exploration into each cluster, integration of external data sources (such as socio-economic indicators or crime rates), and time-based analysis could provide a more comprehensive understanding of underlying factors. It is crucial to recognize that while clustering offers a structured view, interpreting clusters demands domain knowledge and meticulous analysis to avoid over-generalization or erroneous conclusions.

When scrutinizing the primary factors influencing fatal police shootings, age, race, whether the person is armed, and the threat level (whether the person is threatening to attack or actively attacking the police) emerge as the most decisive. However, predicting the non-fatality of incidents proves challenging due to the prevalent recording of predominantly fatal cases in the dataset, warranting caution in drawing accurate predictive conclusions.

4. Appendix A: Method

The data, sourced in CSV format from The Washington Post Data repository, was seamlessly imported into Google Colab for further analysis. This dataset encompassed 8002 fatal incidents, detailing factors such as age, race, armament, police department involvement, and incident location specifics like state and city. To address missing values, we classified them as "undetermined" or assigned median values for numerical attributes like age. Utilizing Python visualization libraries such as Seaborn and Matplotlib, we visualized the distribution of various factors. Additionally, a scatter plot was generated, providing a spatial representation of the United States based on latitude and longitude coordinates, offering insights into incident concentrations across different regions, aligning somewhat with major population centers.

The application of Simple Exponential Smoothing (SES) in our analysis served as a robust time series forecasting method to predict incident counts for the next 12 months. Through a single smoothing parameter, alpha, SES assigned exponentially decreasing weights to past observations, giving precedence to recent incidents. This method proved effective for relatively stable time series data lacking pronounced trends or seasonality.

For clustering based on specific factors, we employed clustering algorithms like K-Means. After preprocessing the data by encoding categorical variables and scaling features, the optimal number of clusters was determined using the "Elbow Method." This approach unveiled distinct clusters within the dataset, grouping incidents based on similarities in factors like gender, race, flee behavior, and threat level. K-Means facilitated a structured framework, revealing inherent patterns and relationships among these factors and providing valuable insights into the diverse dynamics associated with fatal police incidents.

Furthermore, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was leveraged to discern meaningful clusters based on geographical attributes like latitude and longitude. This approach uncovered spatially dense regions, aiding in identifying patterns in the distribution of fatal police incidents across the United States. Visualization of DBSCAN clusters on a USA map provided a comprehensive understanding of spatial dynamics, revealing concentrations in densely populated areas and regions along major highways. This spatial insight serves as a valuable tool for policymakers and law enforcement agencies seeking targeted interventions.

In the feature importance analysis, a RandomForestClassifier was employed, using features like 'id,' 'latitude,' 'longitude,' 'age,' 'armed,' 'race,' 'flee,' 'threat_level,' 'signs_of_mental_illness,' 'body_camera,' 'gender,' and 'is_geocoding_exact.' The results highlighted the significance of factors such as 'id,' 'latitude,' 'longitude,' and 'age,' underscoring the importance of incident location and age. Additionally, features like 'armed,' 'race,' and 'flee' were identified as crucial determinants, shedding light on the impact of weaponry, racial factors, and fleeing behavior on fatal outcomes. This analysis provides valuable insights for policymakers and law enforcement agencies to inform discussions and interventions regarding the identified influential factors.

To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented. Recognizing the potential biases stemming from imbalanced class distributions, SMOTE generated synthetic instances of the minority class, creating a more balanced dataset for model training. This approach mitigated potential biases and enhanced the classifier's ability to generalize to diverse scenarios, contributing to a more comprehensive understanding of factors influencing fatal police incidents.
the Synthetic Minority Over-sampling Technique (SMOTE) from the imbalanced-learn library is utilized to address class imbalance in the training data focused on predicting fatal incidents. SMOTE is applied to the training data to generate synthetic instances of the minority class (non-fatal incidents) and rebalance the dataset. The distribution of the resampled target variable is checked, showcasing the normalized counts of the resampled classes. A Random Forest classifier is then trained on the resampled data, and its performance is assessed on the original test. Evaluation metrics, including accuracy, precision, recall, F1-score, and the confusion matrix, are calculated and stored in the 'evaluation_metrics_fatal_resampled' dictionary. This approach allows for a more comprehensive evaluation of the classifier's performance, considering the effects of oversampling on the model's ability to generalize to both classes.

5. Appendix B: Results

	id	age	longitude	latitude
count	8002.000000	7499.000000	7162.000000	7162.000000
mean	4415.429643	37.209228	-97.040644	36.675719
std	2497.153259	12.979490	16.524975	5.379965
min	3.000000	2.000000	-160.007000	19.498000
25%	2240.250000	27.000000	-112.028250	33.480000
50%	4445.500000	35.000000	-94.315000	36.105000
75%	6579.750000	45.000000	-83.151500	40.026750
max	8696.000000	92.000000	-67.867000	71.301000

Count: There are 8,165 recorded ages. Mean: The average age of individuals is approximately 37.3 years. Std Deviation: The standard deviation is around 13 years, indicating the spread of the age data. Min: The youngest individual was 2 years old. 25% Percentile: 25% of the individuals were 27 years old or younger. Median (50% Percentile): The median age is 35 years. 75% Percentile: 75% of the individuals were 45 years old or younger. Max: The oldest individual was 92 years old.

	Missing Values	Percentage (%)
race	1517	18.957761
flee	966	12.071982
latitude	840	10.497376
longitude	840	10.497376
age	503	6.285929
name	454	5.673582
armed	211	2.636841
gender	31	0.387403

For filling these missing values we used flee strategy and got the output as below:

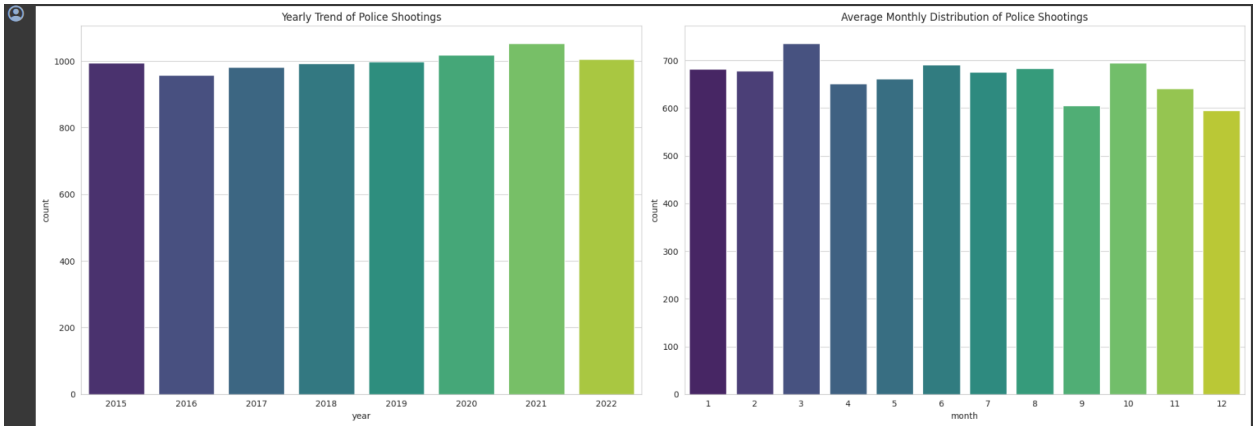
```
race         1517
longitude     840
latitude      840
dtype: int64
```

The age distribution is slightly right-skewed, indicating that younger individuals are more frequently involved in police shootings. The peak occurs around the late 20s to early 30s. Distribution of Gender:

Males dominate the dataset, accounting for the vast majority of police shootings. We also now have an "undetermined" category due to filling in missing values. Top 10 Armed Status Distribution:

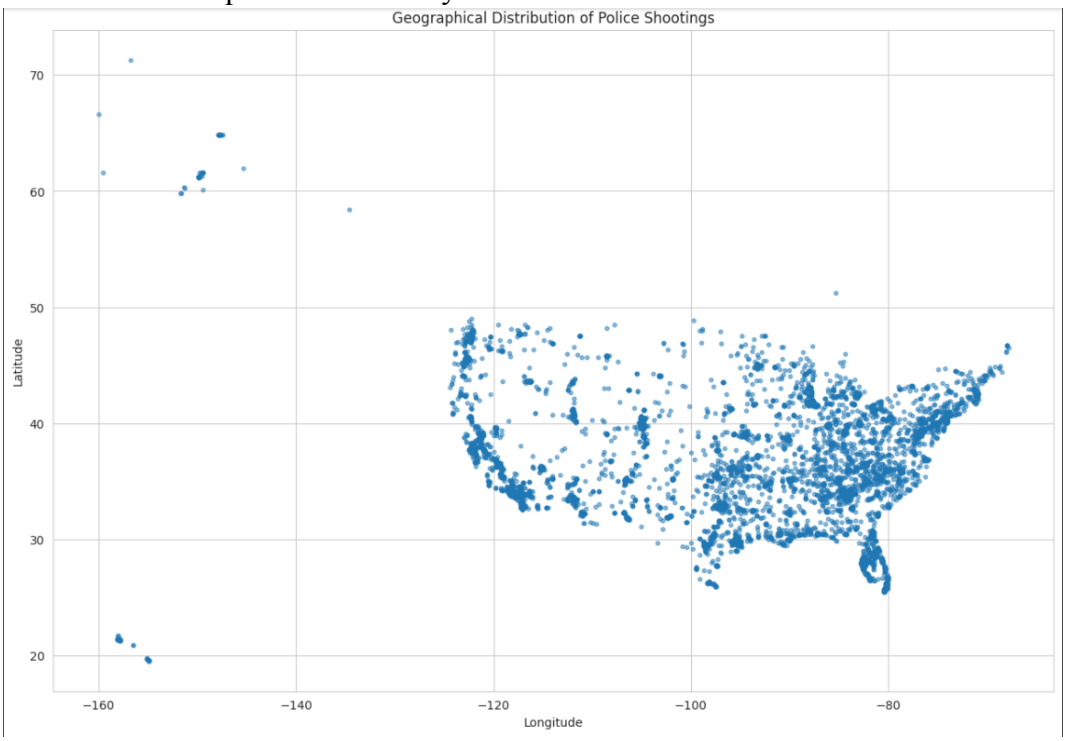
A significant number of individuals were armed with guns. The second most common category is "knife," followed by unarmed individuals. Distribution of Race:

White individuals represent the largest group in the dataset, followed by Black and Hispanic individuals.



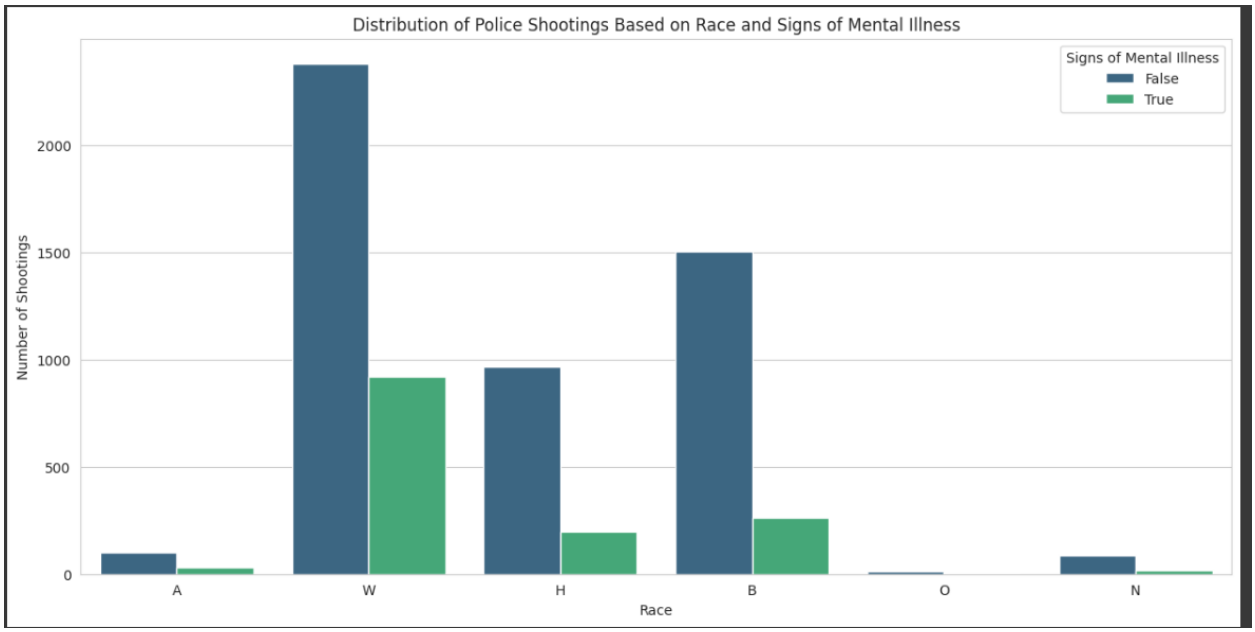
Here's the analysis of police shootings based on the days of the week:

Incidents are fairly evenly distributed across the days of the week. A slight increase is observed on Wednesdays and Fridays, while Sundays tend to have slightly fewer incidents compared to other days.



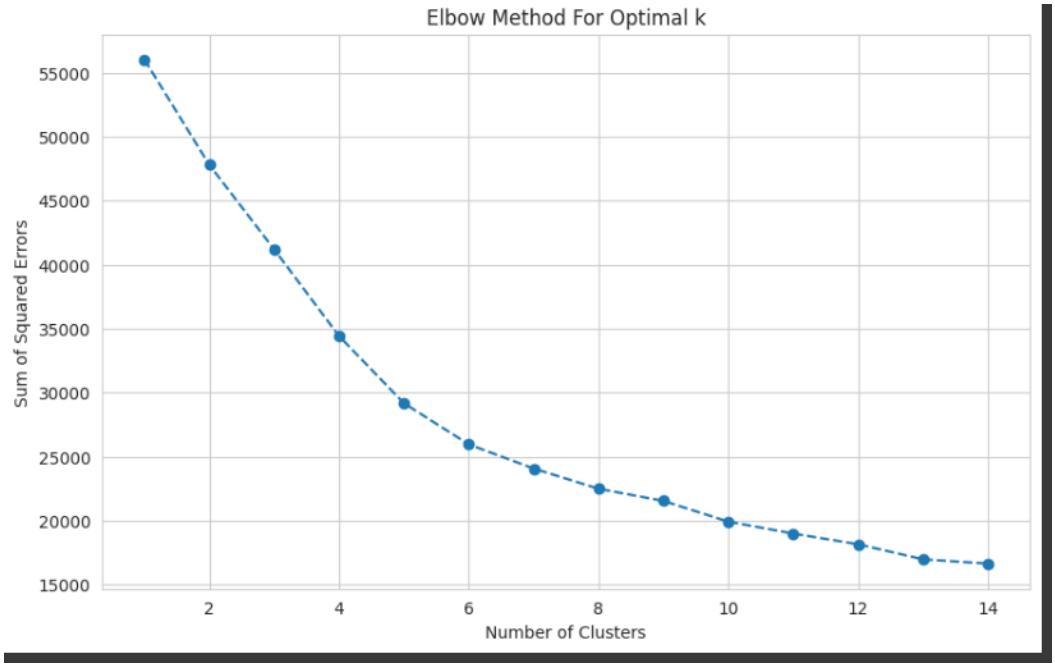
The scatter plot provides a spatial representation of the United States based on the latitude and longitude coordinates of each police shooting incident. We can observe concentrations of incidents in various regions, which somewhat align with major population centers in the U.S.

Relationships and Correlations: We'll investigate the relationship between age, gender, race, and signs of mental illness. Starting with the distribution of age based on the presence or absence of signs of mental illness:



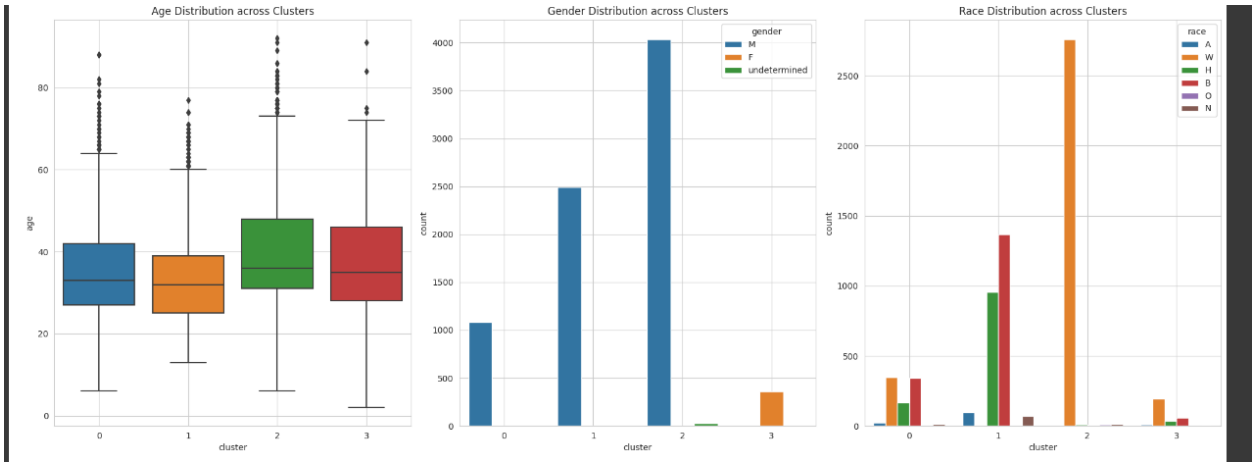
Here's the distribution of police shootings based on race and the presence or absence of signs of mental illness:

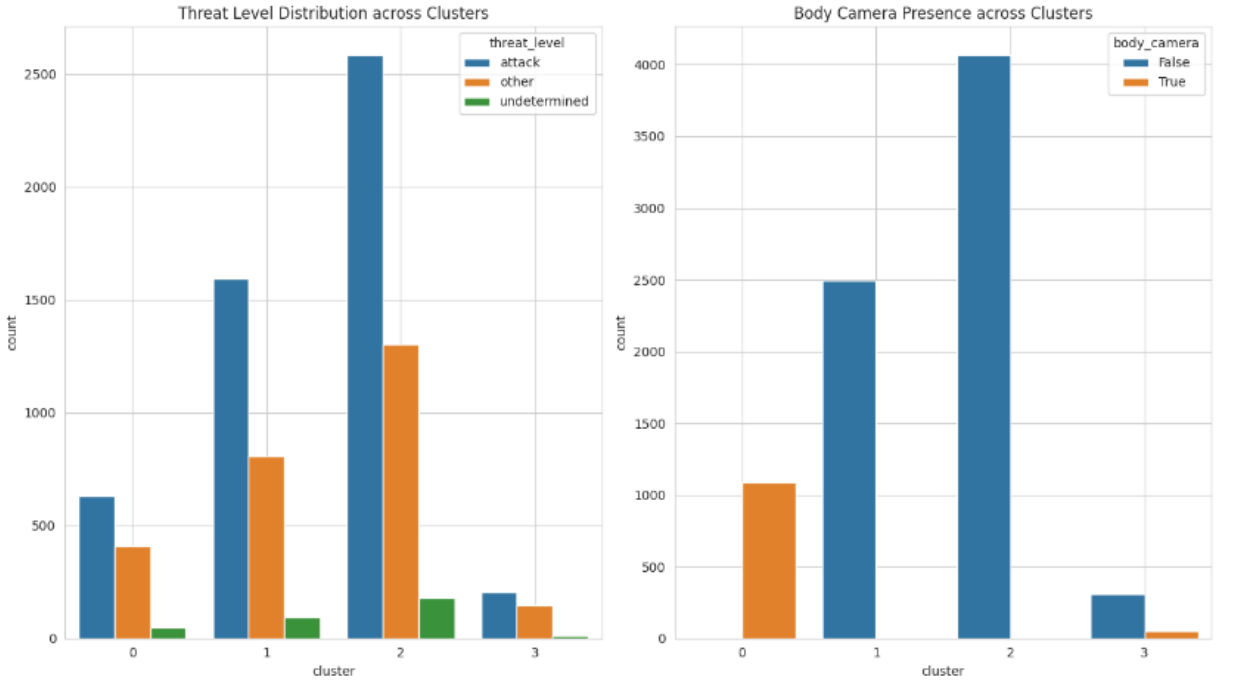
For most racial categories, a larger number of individuals did not show signs of mental illness compared to those who did. The disparity between individuals showing signs of mental illness and those who did not is particularly pronounced in the White and Black categories. This analysis provides insights into the intersection of race and mental health in the context of police shootings.



The Elbow Method graph displays the Sum of Squared Errors (SSE) for a range of cluster numbers. Ideally, we're looking for a point where the reduction in SSE begins to slow

down, indicating a diminishing return in increasing the number of clusters. This point is often referred to as the "elbow."





Age Distribution across Clusters: The boxplots show the distribution of ages in each cluster. While there are some variations in the median age across clusters, the age distributions are relatively similar, with Cluster 1 having a slightly younger age range compared to the others.

Gender Distribution across Clusters: Cluster 3 is predominantly female, while the other clusters are overwhelmingly male.

Race Distribution across Clusters: Cluster 0 and Cluster 2 are predominantly white, while Cluster 1 has a high representation of black individuals.

Threat Level Distribution across Clusters: The threat level is predominantly "attack" across all clusters, with some variations in the "other" and "undetermined" categories.

Body Camera Presence across Clusters: Cluster 0 has a significant number of incidents where body cameras were present, while in the other clusters, body cameras were mostly not present.

Evaluation Metrics:

```
{'Accuracy': 0.9587757651467833,
 'Precision': 0.959375,
 'Recall': 0.9993489583333334,
 'F1-Score': 0.978954081632653,
 'Confusion Matrix': array([[   0,   65],
        [   1, 1535]])}
```

Accuracy: 95.88% This indicates that the model correctly predicted whether an incident was fatal or not for approximately 95.88% of the cases in the test set. Precision: 95.94% Out of all the incidents that the model predicted as fatal, approximately 95.94% were actually fatal. Recall: 99.93% Out of all the actual fatal incidents, the model correctly predicted 99.93% of them. F1-Score: 97.90%. A high F1 score indicates that the model has both good recall and good precision. Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's predictions. True Negative (TN): 0 The number of incidents that were correctly predicted as not fatal. False Positive (FP): 65 The number of incidents that were wrongly predicted as fatal. False Negative (FN): 1 The number of fatal incidents that were wrongly predicted as not fatal. True Positive (TP): 1535 The number of fatal incidents that were correctly predicted.

Predicting the likelihood of an incident being captured on a body camera:

```
{'Accuracy': 0.841995841995842,
 'Precision': 0.1111111111111111,
 'Recall': 0.05454545454545454,
 'F1-Score': 0.07317073170731707,
 'Confusion Matrix': array([[402,  24],
        [ 52,   3]])}
```

6. Appendix C: Code

```
import pandas as pd
# Load the CSV file into a DataFrame
data = pd.read_excel("/content/fatal-police-shootings-data.xls")
# Fitting the Simple Exponential Smoothing model
ses_model = SimpleExpSmoothing(monthly_counts['count']).fit()
# Forecasting for the next 12 months
ses_forecast = ses_model.forecast(steps=12)
# Dates for the forecasted period
forecast_dates_ses = pd.date_range(monthly_counts.index[-1] + pd.DateOffset(months=1), periods=12, freq='M')
# Scaling the features
scaler = StandardScaler()
clustering_data_scaled = scaler.fit_transform(clustering_data)
clustering_data_scaled
from sklearn.cluster import KMeans
# Calculating the Sum of Squared Errors (SSE) for a range of cluster numbers
sse = []
cluster_range = range(1, 15)
for k in cluster_range:
```

```
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(clustering_data_scaled)
    sse.append(kmeans.inertia_)
# Re-encoding categorical variables with newly initialized label encoders
label_encoders_new = {}
for column in ['gender', 'race', 'flee', 'threat_level']:
    le_new = LabelEncoder()
    clustering_data[column] = le_new.fit_transform(clustering_data[column].astype(str))
    label_encoders_new[column] = le_new
 # Calculating characteristics for numerical and categorical features
 numerical_stats = cluster_data[['age']].mean().to_dict()
categorical_stats = {column: cluster_data[column].mode()[0] for column in ['gender', 'race', 'flee',
'threat_level', 'body_camera', 'signs_of_mental_illness']}
# Reducing the dataset size by randomly sampling a fraction of it
sample_data = data.sample(frac=0.3, random_state=42)
# Simplified features and target for predicting fatality
target_fatal = 'manner_of_death'
simplified_features_fatal = ['age', 'gender', 'race']
# Preparing the data
X_fatal_simplified = sample_data[simplified_features_fatal]
y_fatal_simplified = sample_data[target_fatal].apply(lambda x: 1 if x == "shot" else 0)  # 1 for fatal (shot),
0 otherwise

# Encoding categorical variables
for column in X_camera.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    X_camera[column] = le.fit_transform(X_camera[column].astype(str))
# Splitting the data into training and test sets (80% train, 20% test)
X_train_camera, X_test_camera, y_train_camera, y_test_camera = train_test_split(X_camera, y_camera,
test_size=0.2, random_state=42)
# Training a Random Forest classifier
rf_classifier_camera = RandomForestClassifier(random_state=42, n_estimators=100)
rf_classifier_camera.fit(X_train_camera, y_train_camera)
# Predictions on the test set
y_pred_camera = rf_classifier_camera.predict(X_test_camera)
# Reducing the dataset size by randomly sampling a fraction of it for the body camera prediction task
sample_data_camera = data.sample(frac=0.3, random_state=42)
# Encoding categorical variables
for column in X_camera.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    X_camera[column] = le.fit_transform(X_camera[column].astype(str))
# Splitting the data into training and test sets (80% train, 20% test)
X_train_camera, X_test_camera, y_train_camera, y_test_camera = train_test_split(X_camera, y_camera,
test_size=0.2, random_state=42)
# Training a Random Forest classifier
rf_classifier_camera = RandomForestClassifier(random_state=42, n_estimators=100)
rf_classifier_camera.fit(X_train_camera, y_train_camera)
# Predictions on the test set
y_pred_camera = rf_classifier_camera.predict(X_test_camera)
```

Google collab link:
https://colab.research.google.com/drive/1GMjqvA2zg9kLsL2snFv4dyq08ficHhvP?usp=chrome_ntp

***** All of us contributed to the project equally.**