“Spatial and Demographic Analysis of Police Shootings Related to Police Station Locations in the United States"

Co-Authored By:

Aditya Domala - 02096198

Mokesh Balakrishnan - 02126912

Ruksar Lukade - 02137513

Issues:

This data, obtained from the Washington Post, covers police shootings in the United States from 2015 to 2023. The first dataset contains information on police shooting incidents, including the ID, name of the victim, date of the incident, mode of death, armed status, age, gender, race, city, state, signs of mental illness, degree of threat, status of fleeing, availability of body cameras, and geolocation data. The latitude and longitude of US police stations are listed in the second dataset. It offers the geographic coordinates of these stations, which are useful for examining how law enforcement organisations are dispersed and covered across the nation. By examining this information, one can learn more about the features of police shootings, including the demographics of the individuals

The question we are aiming to address:

1. Is there any pattern in the distribution of police stations across different geographic areas, and what disparities might exist?
2. What is the relationship between the geographic distribution of police stations and the accessibility and coverage of law enforcement services across the country?
3. Are there differences in police shootings when age, gender, race, armed status, and signs of mental illness are considered?
4. What patterns can be found in police shootings, and what do these patterns reveal about the dynamics and potential biases in these encounters?
5. How do age, gender, race, armed status, and signs of mental illness affect an individual's likelihood of fleeing during a police encounter?
6. What distinguishes these distinct groups, and what do they tell about distinct profiles or types of police encounters?

Findings:

1. Based on the data provided, we discovered that 8002 events occurred, with victims averaging 37.2 years old.

2. The most prevalent cause of death is a gunshot wound.

3. The whole incident's race distribution is white (50.89%), black (27.23%), Hispanic (17.98%), Asian (1.99%), Native American (1.62%), and others (0.29%).

4. The percentage of males in the overall number of occurrences is 95.51, while the percentage of women is 4.49.

5. We also discovered the top five states with the most incidences, which are California (1143), Texas (732), Florida (509), Arizona (363), and Georiga (306), as well as the top five years, which are 2021 (1053), 2020 (1019), 2022 (1006), 2019 (999), and 2015 (994).

6. There were indicators of mental disorder in 20.88 percent of the occurrences.

7. Police officers' bodycams captured about 14.21 percent of occurrences.

8. The northernmost station is located in Alaska and is located at latitude 71.2921, longitude -156.786.

9. The southernmost station is located in southern US territory, such as Puerto Rico, at latitude 17.9608 and longitude -66.4053.

10. The easternmost station is located in the Caribbean area at latitude 18.3018 and longitude -65.2973.

11. The westernmost station is located in Alaska's Aleutian Islands at latitude 57.125 and longitude 170.285.

Discussions:

Incident Analysis: This would entail mapping and analysing the distribution of fatal police shootings across states and cities in the United States in order to identify areas with higher incidence rates. This analysis can reveal geographical hotspots or patterns.

Demographic Insights: Examining the age, gender, and race of those involved in these shootings may reveal significant patterns or disparities. This knowledge is critical for identifying demographic vulnerabilities and tailoring preventive measures.

Circumstantial Factors: Exploring the context of these shootings, with a particular focus on the victims' mental illness status, whether they were armed, and the perceived threat level. This analysis can assist in comprehending the complexities and causes of such incidents.

Temporal Trends: Examining the time patterns of these incidents to see if certain months or years have higher frequencies. This can reveal the impact of outside factors such as policy changes or social movements.

Body Camera Usage: Discussion of the prevalence and impact of body camera use during these incidents. This is critical in the context of police accountability and the efficacy of such technologies in law enforcement.

Geographical Distribution: Examining and comparing the distribution and density of police stations to regional population density or crime rates. This will reveal details about how law enforcement resources are distributed across the country.

Proximity Analysis: Examining the geographical proximity of police stations to fatal shooting locations. This is critical for understanding law enforcement response times and presence in different areas.

Resource Allocation: Investigating the relationship between the distribution of police stations and resource allocation and access to law enforcement services. This can shed light on regional disparities or coverage gaps in law enforcement.

Appendix A: Method

Data Collection and Preprocessing: We obtained a comprehensive dataset detailing police-involved shootings from the Washington Post, which was presented in Excel. This dataset contains critical information such as each incident's age, gender, ethnicity, and geographic coordinates (latitude and longitude), among other details. We conducted a detailed analysis of this dataset during the preprocessing phase to identify key factors for our research, laying the groundwork for insightful interpretation and findings.

Geospatial Mapping: We focused on the geospatial aspects of police-involved shootings using the same detailed dataset obtained from the Washington Post. The dataset, which was rich in attributes such as age, gender, ethnicity, and precise geographic coordinates (latitude and longitude), allowed for a thorough examination. This analysis enabled us to identify significant trends and patterns, particularly in the spatial distribution of these incidents, adding to the depth and breadth of our research."

Despite this limitation, we created geospatial visualizations in order to identify regional disparities or patterns in police shootings. Our analysis revealed a concentration of incidents along the west and east coasts, with fewer incidents in the center of the country. This observation raised concerns about potential regional disparities, such as population density, law enforcement practices, or demographic characteristics.

Analysis of Age Distribution: We then examined the age distribution of individuals involved in police shooting incidents. To achieve this, we drew the age values from the relevant fields and pre-treated the data to manage incomplete or deviating data. Thereafter, deposit diagrams were formed for all races and racial conglomerates, comprising black, white, and Hispanic people. Numerical data was calculated to comprehend the age grouping of the victims and probable age-associated trends.

Statistical scrutiny: Analyzing data was done to explore potentially dissimilar ages between separate racial cohorts. We did T-tests to discern whether there were notable age variations between whites and blacks or whites and Hispanics. Additionally, an analysis of variance (ANOVA) was used to detect any statistically meaningful sets that shared similar ages.

To assess the recorded mean age differences, Monte Carlo simulations were applied, specifically evaluating white and black ages, as well as white and Hispanic ages. The simulations demonstrated the extent of the age disparities between these racial populations.

Conclusions: In police shooting incidents, white people had a distinctively left-leaning age cohort in comparison with black and Hispanic individuals.

Significant disparities in age were evident between racial groups, as indicated by the results of t-tests. Specifically, there was a notable age difference of seven years between black and white victims. The findings were further supported by Monte Carlo simulations, which provided additional evidence of observed discrepancies in age.

In summary, our study has brought attention to both regional and age disparities in fatal police shootings. These findings underscore the necessity for continuous research and analysis to gain a deeper understanding of the underlying factors that contribute to such inequalities. By emphasizing the significance of ongoing efforts to promote fairness, equality, and transparency in law enforcement encounters, this study contributes to the broader discourse surrounding police shootings.

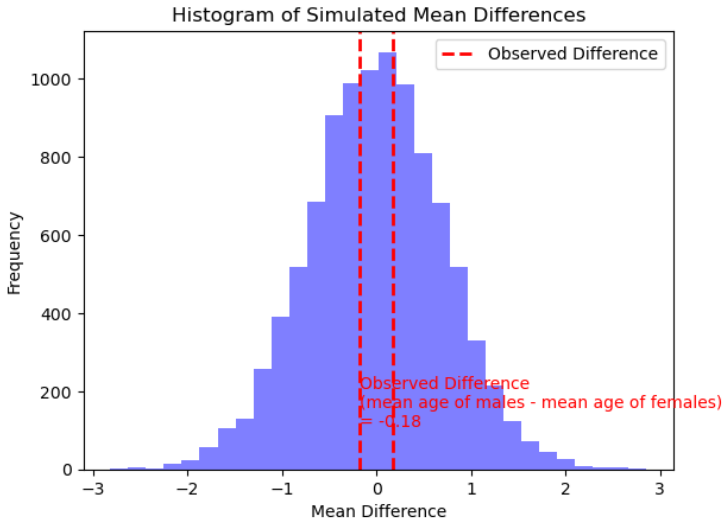
Appendix B: Results

Comparing averages through Monte Carlo simulation

We compared the mean ages of victims based on gender (Males vs Females). We performed the following steps

Dropping rows with missing values in 'age' or 'gender' columns, Formulating our hypothesis, Calculating the observed difference in mean ages, Using Monte Carlo simulation to generate the distribution of mean differences under null hypothesis and calculating the p-value.

```
Out[3]: -0.181979355199823
```



The computed difference in mean age between the male and female participants in police shootings is represented by the observed difference, which is **-0.18**. In particular, the negative sign shows that, on average, the mean age of females in the dataset is somewhat higher than that of males.
Our Monte Carlo test yielded an estimated p-value of roughly **0.402.** The p-value exceeds the standard significance threshold of 0.05. Because of this, we do not reject the null hypothesis and come to the conclusion that there is insufficient data to determine whether the mean age of men and women involved in police altercations differs.

Using Logistic Regression

Here, we have inferred if an individual was fleeing by considering variables such as age, gender, and indications of mental illness.

	precision	recall	f1-score	support
0	0.68	0.88	0.77	845
1	0.57	0.28	0.38	487
accuracy			0.66	1332
macro avg	0.63	0.58	0.57	1332
weighted avg	0.64	0.66	0.62	1332

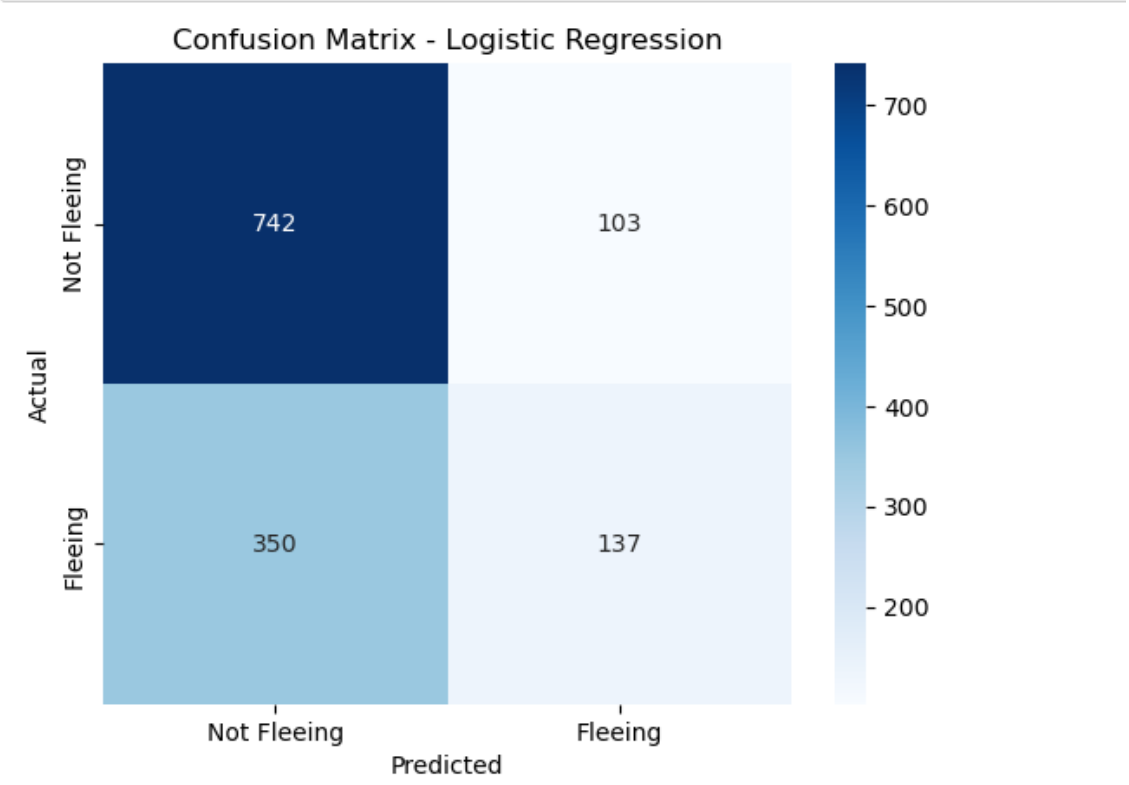
We have examined several metrics, including:

Accuracy: The proportion of accurately predicted positive observations out of all predicted positives. In this case, the accuracy is 0.68 for Non fleeing (0) and 0.57 for fleeing (1).

Sensitivity (recall): The proportion of accurately predicted positive observations out of all observations in the actual class. Here, it is 0.88 for Nonfleeing (0) and 0.28 for fleeing (1).

F1 score: The weighted average of accuracy and sensitivity It aims to strike a balance between accuracy and sensitivity. In this case, it is 0.77 for Non fleeing (0) and 0.38 for fleeing (1).

Support: The number of actual occurrences of the class in the specified dataset. Here, there are 845 instances of Non-Escaping (0) and 487 instances of fleeing (1). The overall accuracy of the model is 0.66 (or 66%), indicating that our model can accurately predict 66% of the cases in our test set.



True Negatives (TN): The top left cell (742) displays the number of occurrences properly predicted as "Not Fleeing". This means that in 742 cases, the model accurately predicted that the individual was not fleeing during a police encounter.

False Positives (FP): The top right cell (103) shows the number of instances where the model mistakenly predicted "fleeing" when the true situation was "Not-Fleeing". These are cases where the model inaccurately anticipated that the individual would flee.

False Negatives (FN): The bottom left cell (350) indicates the number of instances where the model erroneously predicted "Not-Fleeing" when the person was actually "Fleeing". These are scenarios where the model failed to recognize that the person would flee.

True Positives (TP): The bottom right cell (137) reveals the number of occurrences correctly predicted as "Fleeing". This means that the model accurately predicted 137 cases where the person was fleeing.

The intensity of colour represents the number of cases, with darker shades indicating a higher count. This chart is utilised to calculate various performance metrics of the classification model, such as accuracy, precision, recall, and F1 score.

Clustering

We used the K-means algorithm to cluster the events by the categories of 'age','signs of mental illness', and 'flee' in order to analyze them.
Here is the distribution of incidents among the clusters:

```
: 1    3189
  0    3142
  2    1671
  Name: cluster_label, dtype: int64
```

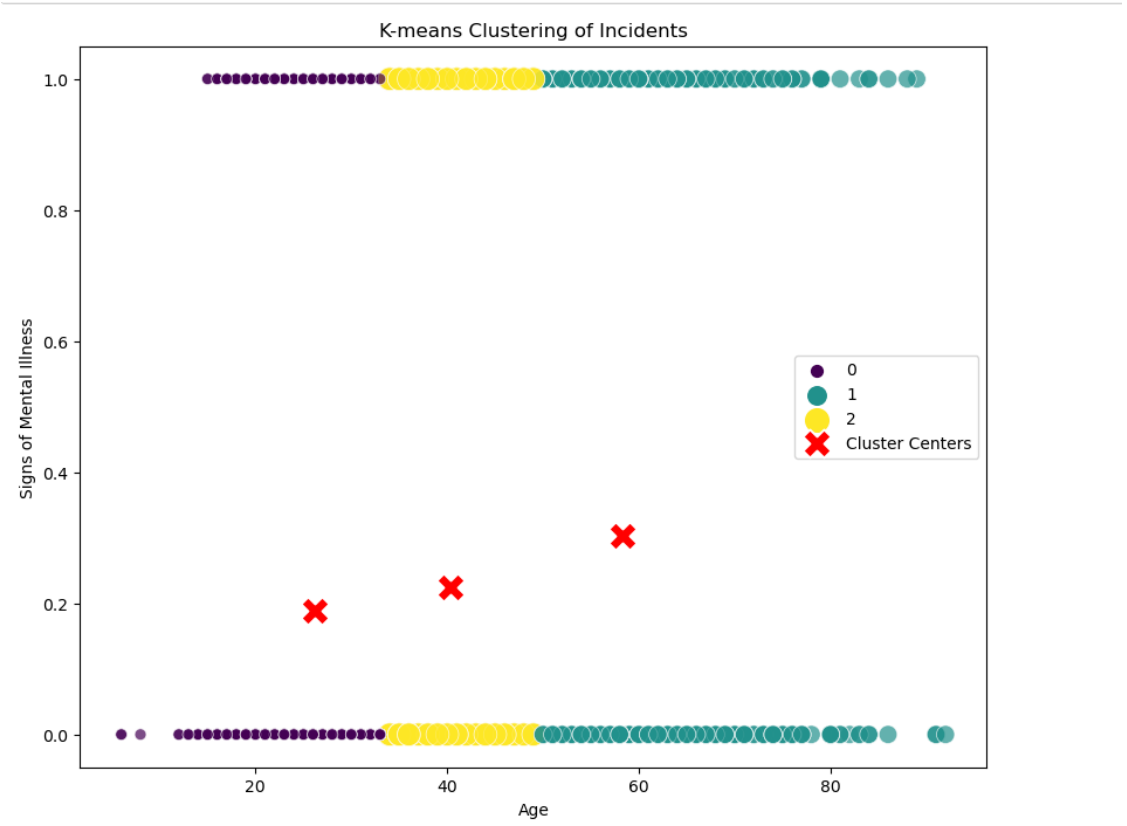
cluster_label	age_mean	signs_of_mental_illness_mean	flee_mean
0	38.292171	0.0	0.000000
1	34.615553	0.0	1.000000
2	39.457810	1.0	0.229204

Here are the average values of each feature within each cluster:

Cluster 0:
Age mean: 38.29
Signs of Mental illness mean: 0.0
Flee mean: 0.0 (Not fleeing)
Cluster 1:
Age mean: 34.62
Signs of Mental illness mean: 0.0
Flee mean: 1.0 (Fleeing)
Cluster 2:
Age mean: 39.46

Signs of Mental illness mean: 1.0
Flee mean: 0.23

Based on these characteristics, the research indicates that the episodes can be divided into three clusters. Incidents in Cluster 0 feature comparatively elderly individuals who do not exhibit symptoms of mental illness or flight. Incidents in Cluster 1 involve comparatively younger people who are fleeing without exhibiting symptoms of mental illness. Incidents in Cluster 2 include those in which the participants are comparatively older, exhibit symptoms of mental illness, and do not flee.



This image is a scatter plot resulting from a K-means clustering analysis. In this plot:

- The x-axis represents the 'age' of individuals involved in the incidents.
- The y-axis is binary and represents the 'Signs of Mental Illness', where 1 indicates presence and 0 indicates absence.
- Data points are colored to represent the cluster each point belongs to after the K-means algorithm has been applied. Three clusters have been identified:
- Cluster 0 (Purple)
- Cluster 1 (Teal)
- Cluster 2 (Yellow)

The red 'X' marks indicate the centroid of each cluster. These are the calculated means of the clustered data points in their respective clusters.

The plot suggests that:

Individuals with no signs of mental illness are spread across all ages (most data points are at y = 0).

The cluster of individuals showing signs of mental illness (y = 1) is relatively small compared to the other clusters.

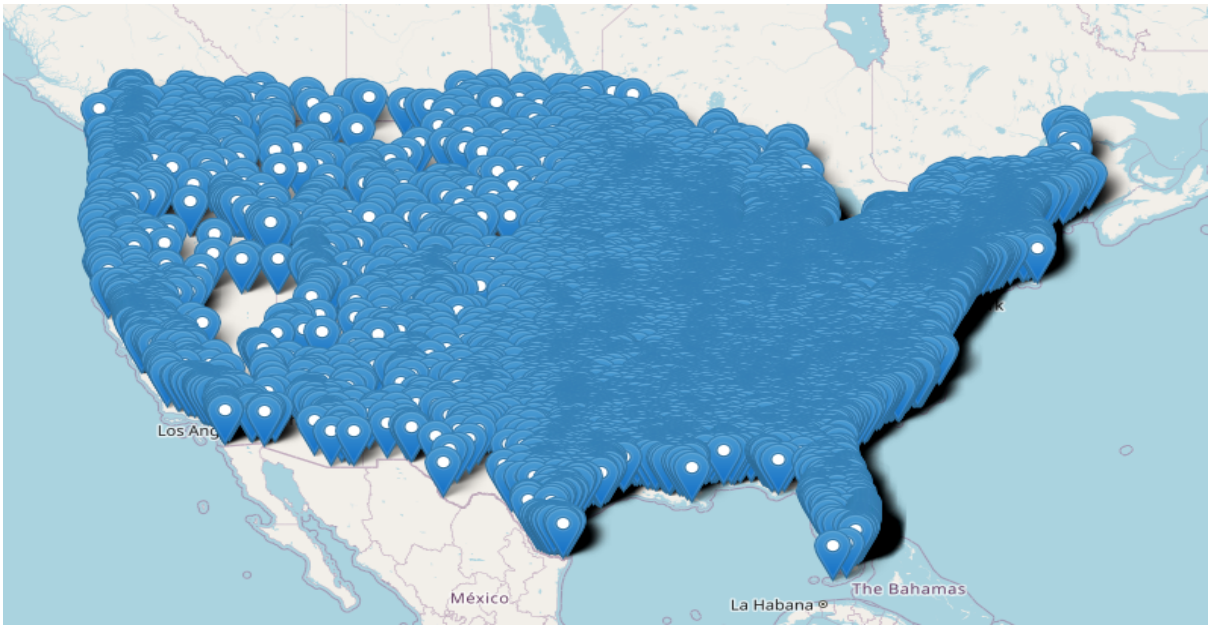
There is a concentration of data points (and thus incidents) at lower ages, indicating that younger individuals are more commonly involved in these incidents, based on this dataset.

The cluster centroids suggest that the average age of individuals in each cluster varies, with one cluster (Cluster 1) having a centroid at a higher age than the others.

This clustering analysis can be useful for identifying patterns and correlations within the data that might warrant further investigation or could inform policy decisions.

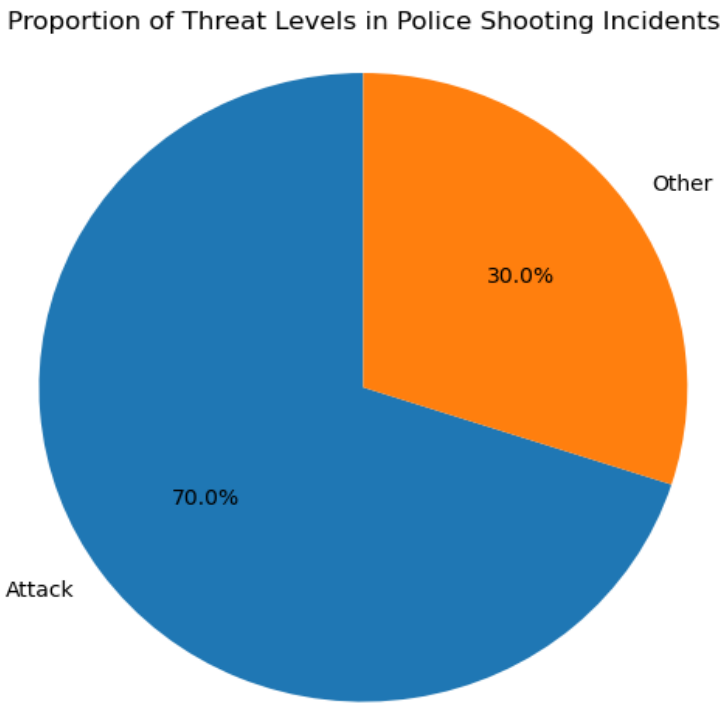
Geographic distribution of police stations:

We have analyzed the concentration of police stations in different areas and identified any geographic patterns or disparities. This helps in understanding the accessibility and coverage of law enforcement services across the country.



Pie Chart - Threat Level:

We constructed a pie chart to display the proportion of different threat levels (e.g., attack, other) in police shooting incidents. This can give a sense of the overall distribution of threat levels during these encounters.



This chart displays the distribution of threat levels reported in police shooting incidents. The chart is divided into two segments:

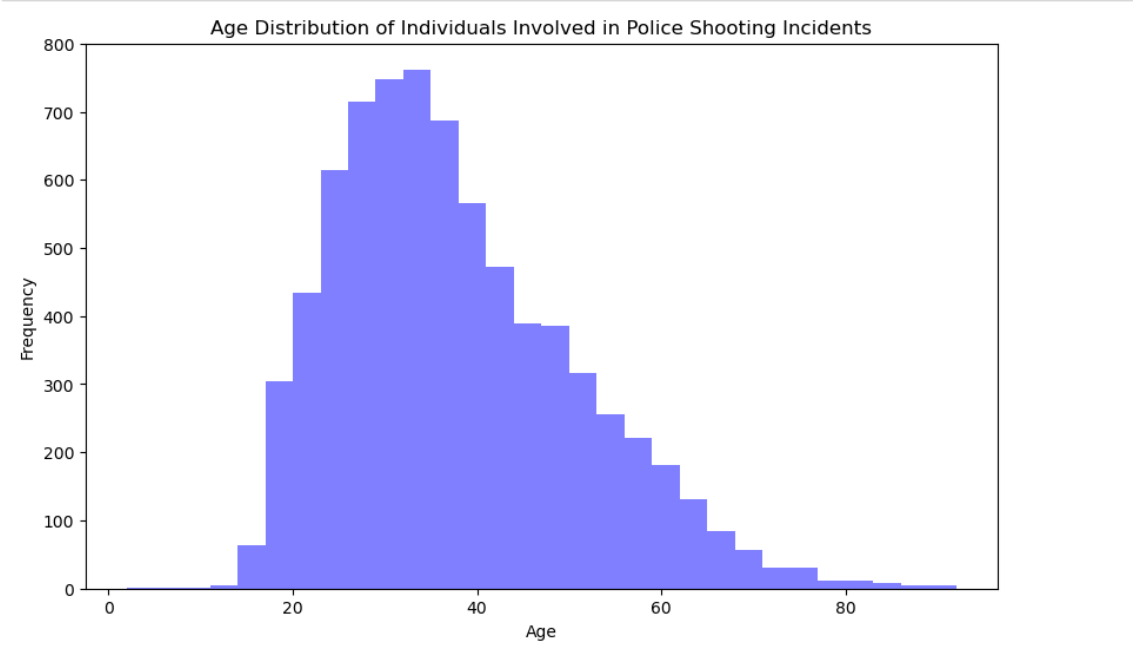
Attack (70%): This large segment of the pie chart represents incidents where the threat level was classified as an 'Attack'. It covers the majority of the chart, indicating that in 70% of the incidents, the individuals were perceived to be attacking or posing an immediate threat when they were shot.

Other (30%): The smaller segment represents all other threat levels combined into a single category. This could include individuals who were classified as posing a lesser threat or no threat at all at the time of the shooting.

The chart provides a visual representation of the relative frequencies of these two categories, highlighting that a significant majority of the incidents are reported under the 'Attack' category. This kind of visualization is helpful for quickly grasping the distribution of categorical data.

Histogram - Age Distribution:

We have generated a histogram to visualize the age distribution of individuals involved in police shooting incidents. This can reveal insights about the age demographics and potential age-related patterns.



This histogram provides a visual representation of the age distribution of individuals involved in police shooting incidents.

Most common age range: The histogram shows the most frequent age or age range at which individuals were involved in police shootings. This is indicated by the height of the bars; the tallest bar signifies the age range with the highest frequency of incidents.

Spread of ages: The range of the bars from left to right indicates the spread of ages among the individuals involved in these incidents. A wider spread suggests a greater age range among the individuals.

Shape of the distribution: The overall shape of the histogram can show whether the distribution of ages is symmetrical, skewed to the left (indicating a younger age profile), or skewed to the right (indicating an older age profile).

Outliers: If there are bars that are significantly separated from the rest of the distribution, this could indicate outliers — ages that are unusually low or high relative to the typical age range of individuals involved in these incidents.

This histogram is a useful tool for quickly assessing the age characteristics of individuals in police shooting incidents and can be critical for identifying whether certain age groups are disproportionately represented.

Appendix C: code

To import data:

In [1]:
```
# Importing the pandas library
import pandas as pd

# Define the file paths
file_path_1 = 'Latitude-longitude of US police stations.csv'
file_path_2 = 'fatal-police-shootings-data.csv'

# Load the files into pandas DataFrames
df1 = pd.read_csv(file_path_1)
df2 = pd.read_csv(file_path_2)

# Display the first few rows of each DataFrame
print(df1.head())
print(df2.head())
```

Monte carlo testing:

In [17]:
```
# Drop rows with missing 'age' or 'gender'
df_mc = df2.dropna(subset=['age', 'gender'])

# Formulate the hypotheses
# Null hypothesis (H0): The mean age of males and females who were involved in police shootings is the same.
# Alternative hypothesis (H1): The mean age of males and females who were involved in police shootings is different.

# Calculate the observed difference in mean ages
obs_diff = df_mc[df_mc['gender'] == 'M']['age'].mean() - df_mc[df_mc['gender'] == 'F']['age'].mean()
obs_diff
```

Out[17]:
```
-0.1781979355199823
```

In [4]:
```
import numpy as np

def monte_carlo_test(group1, group2, num_samples=10000):
    """
    Performing a Monte Carlo test to compare means of two groups.
    """

    # Concatenate the two arrays
    data = np.concatenate([group1, group2])

    # Initialize an empty array to hold our simulated differences
    simulated_diffs = np.empty(num_samples)

    # Run simulations
    for i in range(num_samples):
        # Shuffle the data
        np.random.shuffle(data)

        # Split into two groups
        simulated_group1 = data[:len(group1)]
        simulated_group2 = data[len(group1):]

        # Compute the difference in means
        simulated_diffs[i] = simulated_group1.mean() - simulated_group2.mean()

    return simulated_diffs


# Groups for Monte Carlo Test
group1 = df_mc[df_mc['gender'] == 'M']['age'].values
group2 = df_mc[df_mc['gender'] == 'F']['age'].values

# Run the Monte Carlo Test
simulated_diffs = monte_carlo_test(group1, group2)

# Estimate the p-value
p_value = np.sum(simulated_diffs >= np.abs(obs_diff))/len(simulated_diffs)
p_value
```

Out[4]:
```
0.4014
```

Logistic Regression:

In [5]:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.preprocessing import LabelEncoder
from sklearn import metrics

# Drop rows with missing 'age', 'gender', 'signs_of_mental_illness' or 'flee'
df_lr = df2.dropna(subset=['age', 'gender', 'signs_of_mental_illness', 'flee'])

# Encode 'gender' column numerically
le = LabelEncoder()
df_lr['gender'] = le.fit_transform(df_lr['gender'])

# Encode 'flee' column where 'Not fleeing' is 0 and 'Fleeing' is 1
df_lr['flee'] = df_lr['flee'].apply(lambda x: 0 if x == 'Not fleeing' else 1)

# Defining X (features) and y (target)
X = df_lr[['age', 'gender', 'signs_of_mental_illness']]
y = df_lr['flee']

# Split data into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=1)

# Create a logistic regression model
model = LogisticRegression()

# Train the model using the training sets
model.fit(X_train, y_train)

# Predict Output
y_pred = model.predict(X_test)

# Print the classification report
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.68      0.88      0.77       845
           1       0.57      0.28      0.38       487

    accuracy                           0.66      1332
   macro avg       0.63      0.58      0.57      1332
weighted avg       0.64      0.66      0.62      1332
```

Clustering:

Let's start by preprocessing the data.

In [6]:

```
# Fill missing values in 'age' column with median age
df_cluster = df2.copy()
df_cluster['age'].fillna(df_cluster['age'].median(), inplace=True)

# Binarize 'flee' column where 'Not fleeing' is 0 and 'Fleeing' is 1
df_cluster['flee'] = df_cluster['flee'].apply(lambda x: 0 if x == 'Not fleeing' else 1)

# Select relevant columns for clustering
cluster_data = df_cluster[['age', 'signs_of_mental_illness', 'flee']]

# Scale the features
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
cluster_data_scaled = scaler.fit_transform(cluster_data)

# Perform K-means clustering
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, random_state=1)
kmeans.fit(cluster_data_scaled)

# Assign cluster labels to the incidents
df_cluster['cluster_label'] = kmeans.labels_

# Display the clusters distribution
df_cluster['cluster_label'].value_counts()
```

Out[6]:

```
1    3189
0    3142
2    1671
Name: cluster_label, dtype: int64
```

In [7]:
```
# Calculate the mean values of each feature within each cluster
cluster_summary = df_cluster.groupby('cluster_label').agg(
    age_mean=('age', 'mean'),
    signs_of_mental_illness_mean=('signs_of_mental_illness', 'mean'),
    flee_mean=('flee', 'mean')
)

cluster_summary
```

Out[7]:

cluster_label	age_mean	signs_of_mental_illness_mean	flee_mean
0	38.292171	0.0	0.000000
1	34.615553	0.0	1.000000
2	39.457810	1.0	0.229204

Let's start by preprocessing the data.

In [6]:
```
# Fill missing values in 'age' column with median age
df_cluster = df2.copy()
df_cluster['age'].fillna(df_cluster['age'].median(), inplace=True)

# Binarize 'flee' column where 'Not fleeing' is 0 and 'Fleeing' is 1
df_cluster['flee'] = df_cluster['flee'].apply(lambda x: 0 if x == 'Not fleeing' else 1)

# Select relevant columns for clustering
cluster_data = df_cluster[['age', 'signs_of_mental_illness', 'flee']]

# Scale the features
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
cluster_data_scaled = scaler.fit_transform(cluster_data)

# Perform K-means clustering
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, random_state=1)
kmeans.fit(cluster_data_scaled)

# Assign cluster labels to the incidents
df_cluster['cluster_label'] = kmeans.labels_

# Display the clusters distribution
df_cluster['cluster_label'].value_counts()
```

Out[6]:
```
1    3189
0    3142
2    1671
Name: cluster_label, dtype: int64
```

In [7]:
```
# Calculate the mean values of each feature within each cluster
cluster_summary = df_cluster.groupby('cluster_label').agg(
    age_mean=('age', 'mean'),
    signs_of_mental_illness_mean=('signs_of_mental_illness', 'mean'),
    flee_mean=('flee', 'mean')
)

cluster_summary
```

Out[7]:

cluster_label	age_mean	signs_of_mental_illness_mean	flee_mean
0	38.292171	0.0	0.000000
1	34.615553	0.0	1.000000
2	39.457810	1.0	0.229204

Contribution Statement for the Project Report:

As members of the project team, we hereby declare our individual contributions to the project's implementation:

Aditya Domala:

My main objectives in this project were to investigate and analyze the dataset, with a particular focus on threat kinds, weapon roles, threat demographics, and the rationale of shootings. In the beginning, I calculated the range of the dataset, verified that no values were missing, and estimated its size. I looked at age, race, gender, mental health, and age ranges in my thorough demographic research. I assist in preprocessing the datasets, performing statistical tests and modeling techniques, generating visualizations, and providing insights based on the analysis.

Mokesh Balakrishnan:

Mokesh was primarily in charge of documenting and synthesising our findings for the final report. He made certain that the report was thorough and well-structured, and that our findings were presented in an understandable and insightful manner. His contributions aided in the reconciliation of our technical findings and their presentation.

Ruksar Lukade:

Ruksar was in charge of gathering the resources and references needed to support our project. During the report-writing phase, she collaborated closely with Mokesh, providing valuable insights and ensuring that all aspects of our analysis were fully covered. His dual role ensured that our project was well-informed and well-documented.

All members actively participate in discussion, planning, and decision-making, ensuring that the project is a collaborative effort throughout. We remain committed to our collaborative efforts and the final report that has been presented.

Aditya Domala - 02096198

Mokesh Balakrishnan - 02126912

Ruksar Lukade - 0213751