"From Data to Insights: A Data-Driven Exploration of Fatal Police Shootings in the U.S."

Co-authors

Amith Ramaswamy, Chiruvanur Ramesh Babu Sai Ruchitha Babu, Nikhil Premachandra rao, Prajwal Sreeram Vasanth Kumar

A Data-Driven Exploration of Fatal Police Shootings in the U.S.

THE ISSUES:

Within the field of data science and advanced statistics, our attention is drawn to the major issue of fatal police shootings in the United States. By utilizing the information obtained from the Washington Post, our research is ready to discover knowledge. Our goal is to provide detailed, data-driven knowledge of this crucial problem in society that goes beyond traditional news headlines by analyzing complex trends and identifying significant causes.

The dataset in question contains a wide range of features, including important attributes like gender, armed, race, state, body camera, and signs of mental illness.

We address the questions:

- What are the trends in police shootings within California?
- How can we understand regional trends contributing to the target?
- The probability of a Male being shot in different races.
- Was there any bias when there was a body cam on any race?

THE FINDINGS:

We looked at the information from the Washington Post about when police use deadly force in the U.S., specifically in California. There are some interesting things we can see when we look at the details.

First off, we saw differences in how this happens based on things like age, where it happens, and the relationship between the police and the community. It's important to pay attention to what's going on locally because the rules, the people, and the places all play a role in this.

When we look deeper into the data, we see that different racial groups are affected at different rates, and we're also checking out if there's a difference between guys and girls in these situations. This could tell us if there are problems with how the police are doing things or if there are unfair biases happening.

One interesting discovery is about body cameras. When police officers have these cameras on, it seems like there's less chance of unfair treatment. But we have to be careful – just because there's a connection doesn't mean one causes the other. We might need more research to be sure and to understand what else could be going on.

In a nutshell, this study shows us that looking closely at what's happening in different areas, considering who's involved, and even having body cameras can make a difference in how these situations play out. But we still have more to learn to really understand why things happen the way they do.

DISCUSSION:

We took a close look at how and when the police use deadly force in California, and found some interesting patterns. It seems that certain factors like age, location, and the relationship between the police and the community play a role in these situations.

Understanding what's happening locally is crucial because it affects how these incidents unfold.As we look deeper into the data, I noticed disparities among different racial groups and possibly between genders. This raises questions about whether there are issues with how the police operate or if unfair biases are at play.

One intriguing finding is related to body cameras worn by police officers. It appears that when they use these cameras, there's a lower likelihood of unfair treatment. However, it's important to be cautious – just because there's a connection doesn't mean one thing causes the other. We might need more research to confirm this and to explore other factors that could be influencing the situation.

To Conclude, study suggests that examining specific details in different areas, taking into account who is involved, and even having police wear body cameras can influence how these incidents unfold.

Author Contributions:

	Amith Ramaswamy	Sai Ruchitha Babu	Nikhil Premachandraro	Prajwal S V
Data cleaning	20%	35%	15%	30%
Analyzing	30%	15%	20%	35%
Coding	35%	15%	35%	15%
Visualizing	25%	15%	35%	25%
Report Writing	20%	30%	25%	25%

Appendix A: Method

Data collection:

We accessed the entire police shootings database provided by The Washington Post, available at the following link. This database is useful since it carefully documents fatal police shootings in the United States. The Washington Post carefully selects and updates this material, which comes from various sources, including local news reports, police enforcement websites, and independent databases. Our study acquires a strong foundation of data from which to undertake major analysis and gain insights into the complex dynamics surrounding fatal police shootings by employing this dependable and often updated database.

Variable creation:

Several important variables derived from The Washington Post's police shootings database were examined in our investigation. These variables include a wide range of parameters and demographic information which includes the age, race, and gender of those involved in fatal police shootings. Geographic information such as location and state were also taken into account to comprehend spatial patterns. The use of body cams during events was investigated as an additional factor impacting outcomes. The fact that the shooters were armed offered information on the circumstances underlying the shootings. By carefully integrating these elements, we find trends and understand the complex dynamics present in fatal police shootings in the United States.

Data Cleaning:

- Missing values were handled using different statistical techniques.
- The age column was handled using the KNN imputation method.
- Analysis by city and state was used to replace the missing data in the race column.
- The missing values in the latitude and longitude columns were successfully handled by grouping them by city and state. However, after doing this, some rows were removed since we were unable to replace the columns efficiently.

Analytic methods:

Filling in Missing Data (KNN Imputation):

Our data had some gaps, like missing puzzle pieces. To fix this, we used a method called K-Nearest Neighbors (KNN) imputation. Imagine you're missing a puzzle piece, so you look at the ones nearby to guess what should be there. KNN does something similar; it fills in missing values by checking nearby incidents in our data. This way, we made sure our data was complete and more accurate.

Chi-square Test:

We wondered if a person's background, like their race, is connected to police shootings. To find out, we used a tool called the Chi-square test. It's like being a detective, trying to see if there's a link between what people prefer and where they come from. The Chi-square test helped us figure out if there's a significant link between factors like race and the occurrence of police shootings. It helped us find any patterns or biases based on these backgrounds.

ANOVA Test:

We wanted to know which factors play a big role in decisions leading to police incidents. So, we used something called Analysis of Variance (ANOVA). Think of it like comparing different groups, similar to checking if groups of friends with different habits spend their weekends differently. ANOVA helped us pinpoint which factors significantly influenced the creation of decision trees. This was important in understanding the key elements shaping police shootings.

Bar Graphs:

Numbers and trends can be tricky, so we used simple bar graphs to tell a story visually. These graphs showed the number of police shooting incidents over time for different racial groups. It's like drawing a timeline and marking incidents for each group. We also used bar graphs to show what factors influenced predicting a person's race. These visuals made our data easier to understand, helping us see patterns and trends at a glance.

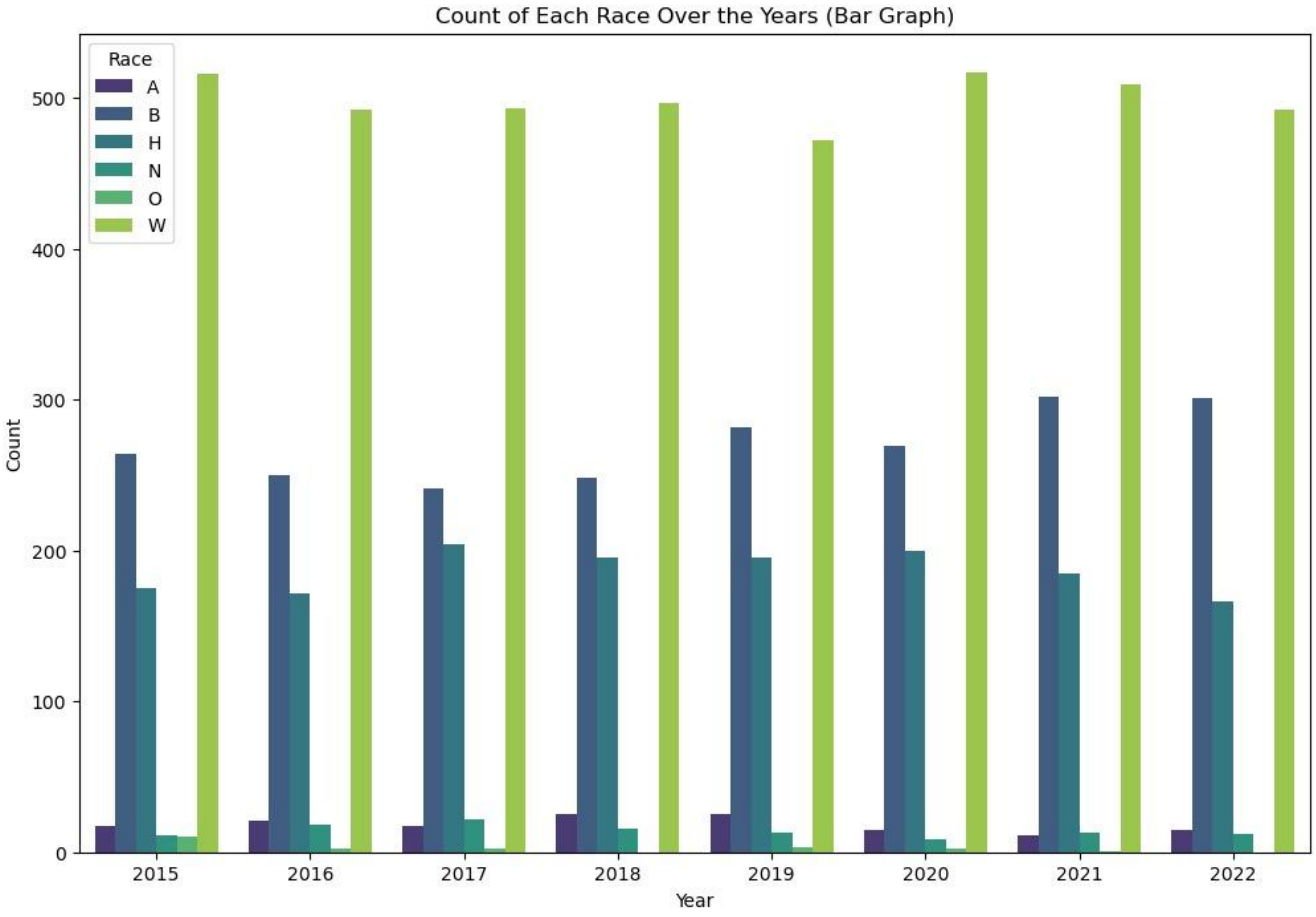
Decision Tree Model:

we wanted to predict outcomes based on important factors. That's where our decision tree model comes in. Imagine it as a roadmap guiding decisions based on conditions. In our case, it helped us understand how different factors link and contribute to the likelihood of police incidents. It's like saying, "If this happens, then the next step might be that." This model aided us in making decisions and policies.

Bayes Rule:

In the world of probability, we wanted to figure out the chances of a male being involved in a police shooting across different racial groups. So, we used Bayes' theorem, a way to calculate probabilities based on what we already know. Picture it like playing a game and figuring out the likelihood of winning under different circumstances. Bayes Rule helped us calculate the chances of a male being shot considering different races. It's using logic to understand the probability of police shootings based on both gender and race, giving us more detailed insights.These methods were like our tools for exploring the complex world of police shootings. Each played a special role, from completing missing info to showing trends and patterns, and from understanding connections to making predictions. Together, they formed the core of our project, helping us understand the intricate factors influencing fatal police incidents.

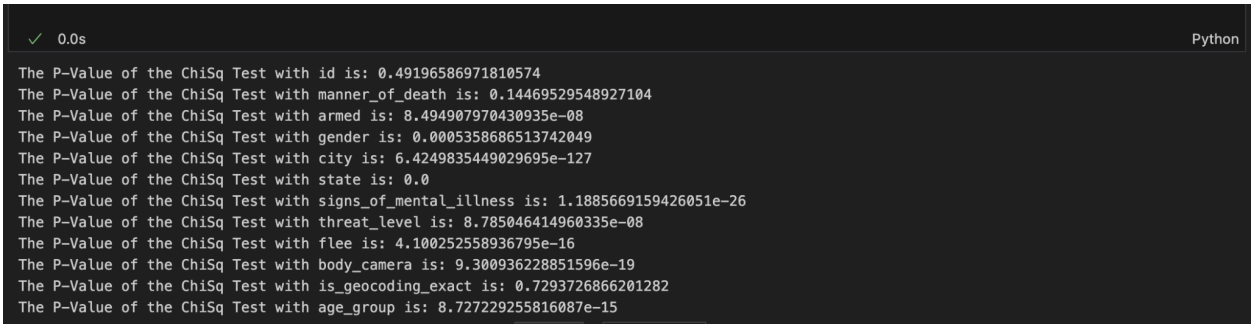
Appendix B: Results



Visualizing the trend in the number of deaths of people over each race over the period from 2015 to 2022 using a bar graph: *from the above graph we can conclude that White and Black are the two races which are high in number.*

```
##### ANOVA Results #####

longitude is correlated with encoded_Race | P-Value: 1.3000106498303888e-22
latitude is correlated with encoded_Race | P-Value: 2.1464370891759254e-20
age is correlated with encoded_Race | P-Value: 2.7161890567760563e-60

['longitude', 'latitude', 'age']
```

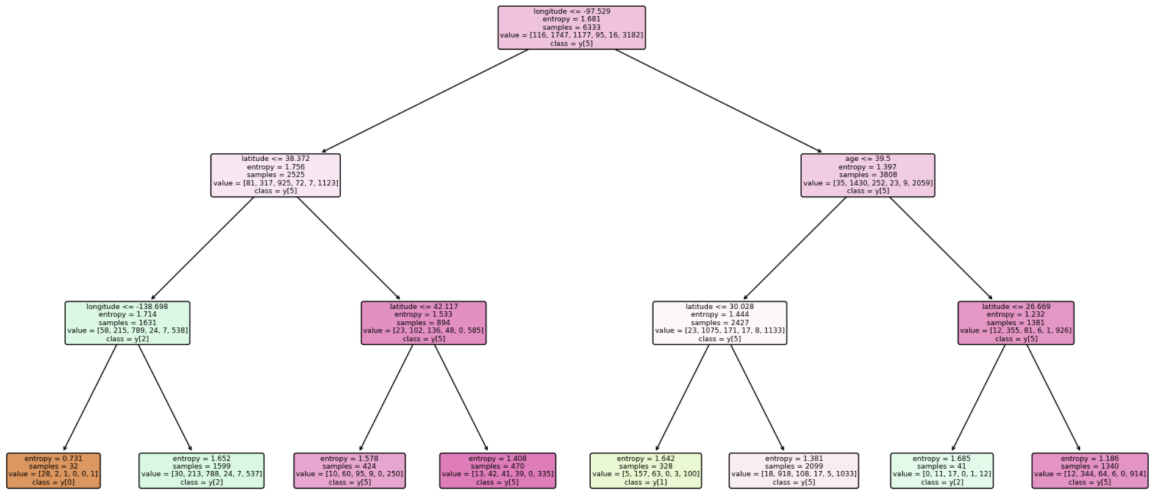
ANOVA test: *we can see that latitude, longitude, and age are significant to the target column race.*



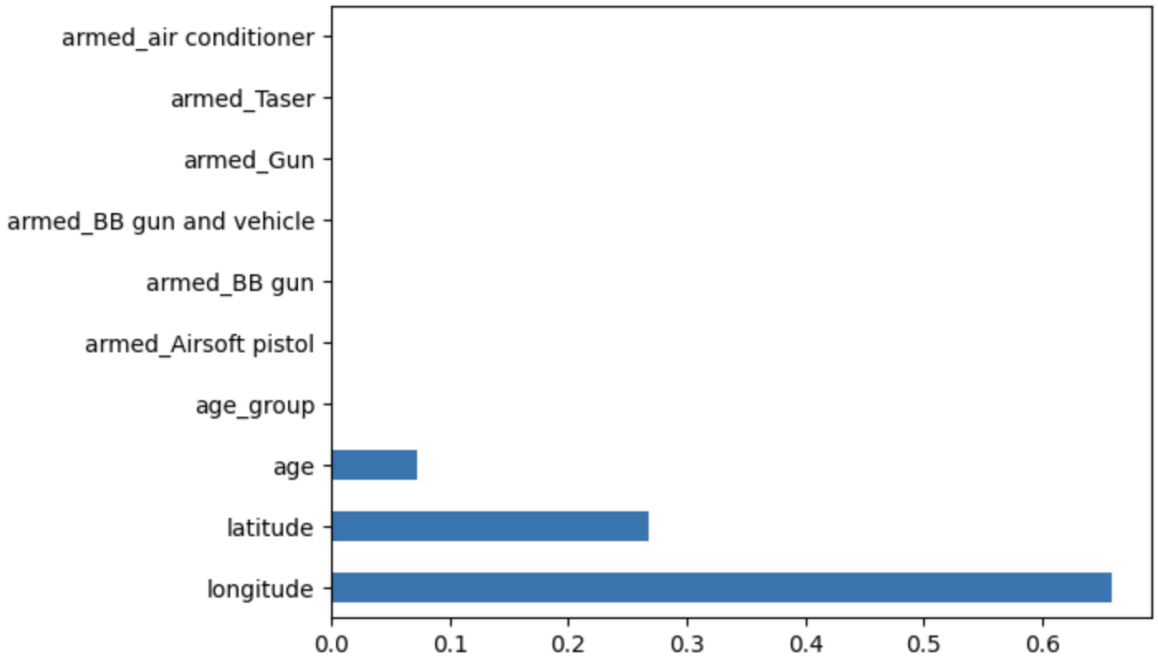
Chi-Square test: *We can consider columns that have a p-value less than a significant value of 0.05 on the target value which is "race"*

```
DecisionTreeClassifier(criterion='entropy', max_depth=3)
              precision    recall  f1-score   support

           0       0.50      0.07      0.12        30
           1       0.43      0.07      0.12       410
           2       0.51      0.70      0.59       315
           3       0.00      0.00      0.00        19
           4       0.00      0.00      0.00         4
           5       0.58      0.78      0.67       806

    accuracy                           0.56      1584
   macro avg       0.34      0.27      0.25      1584
weighted avg       0.52      0.56      0.49      1584

[[  2   0  12   0   0  16]
 [  0  29  49   0   0 332]
 [  1  10 220   0   0  84]
 [  0   0   4   0   0  15]
 [  0   0   2   0   0   2]
 [  1  28 148   0   0 629]]
```

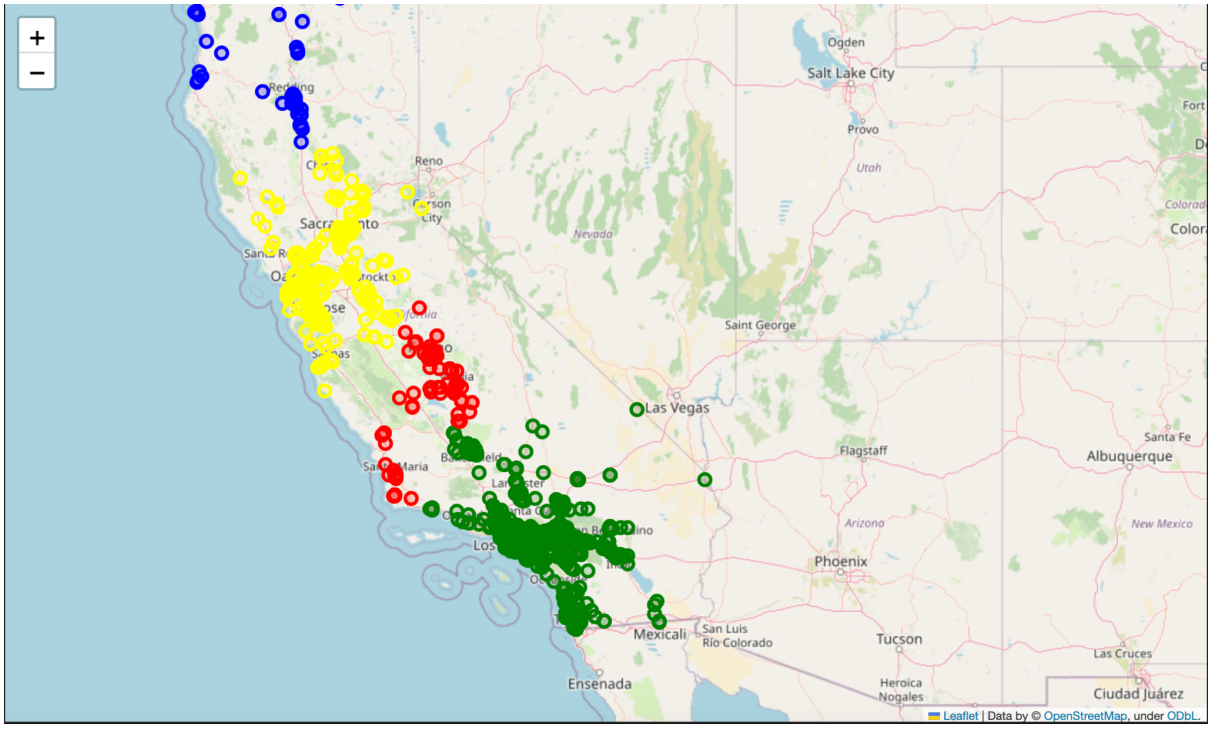
Decision Tree Classifier: *produced an overall accuracy of 56% on the dataset.*



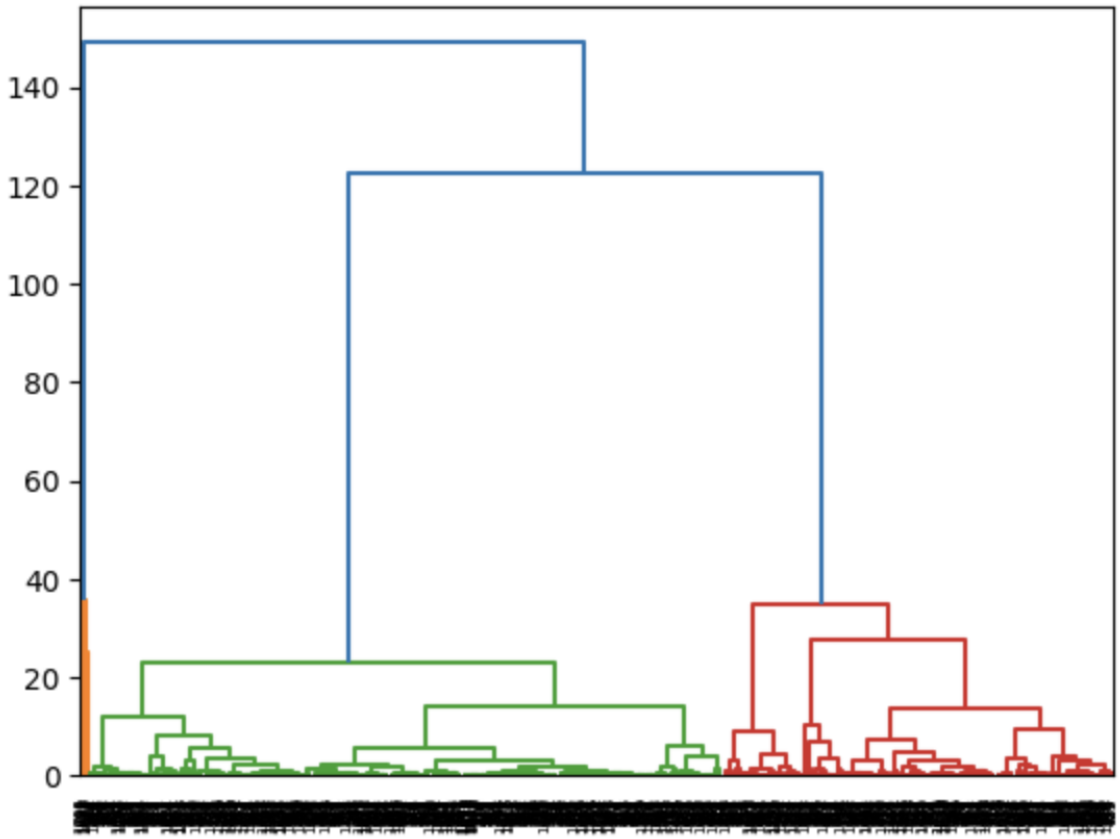
Visualizing a decision tree in the graph: *which breaks down complex choices into clear, understandable branches and outcomes*



Top 10 factors: *that affect predicting the race of a particular person visualized in a bar graph, ”latitude, longitude, age” has the lion’s share.*



Visualizing clusters: *in the state of California based on latitude and longitude, It is evident from the picture that we see distinct clusters.*

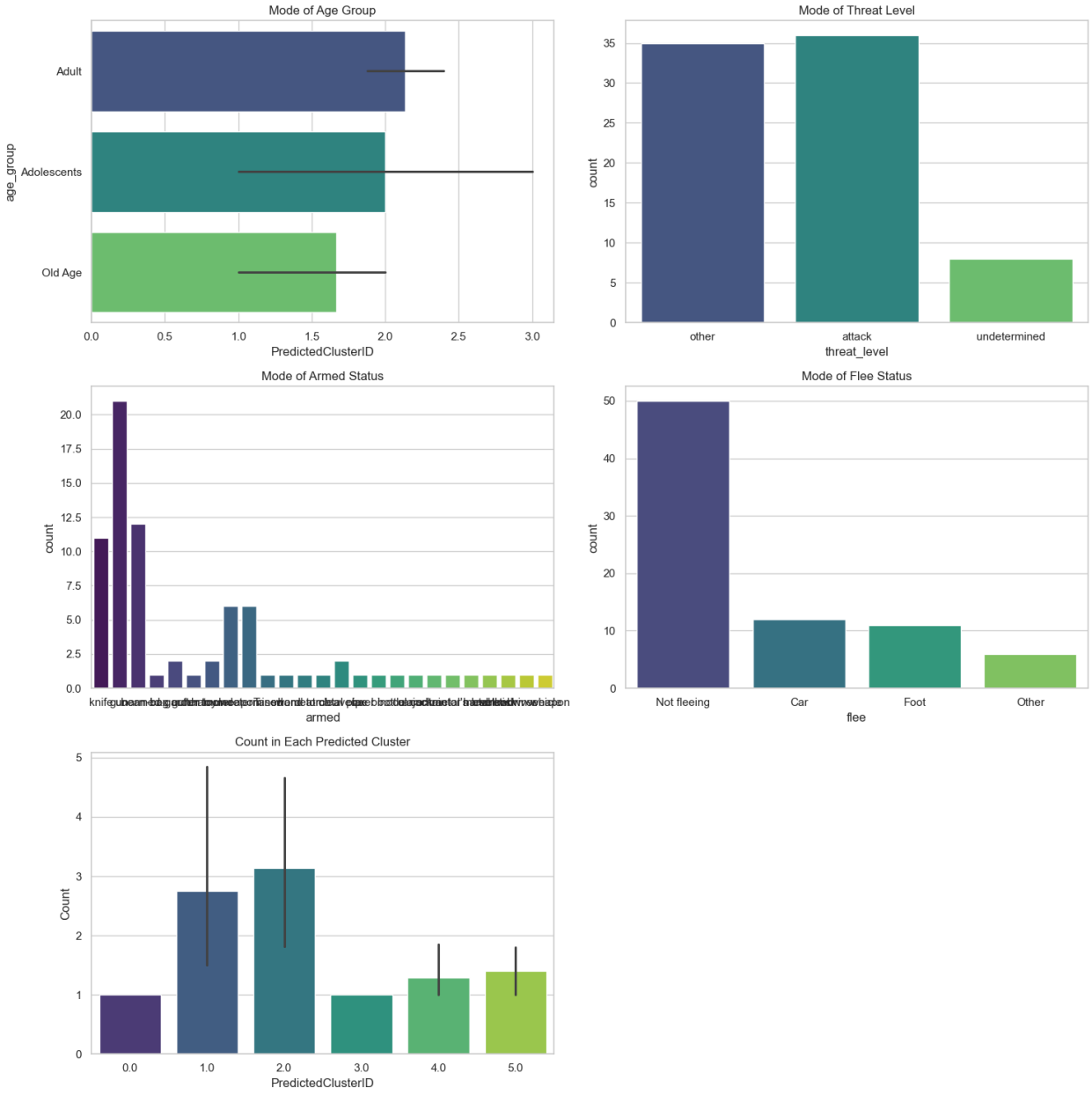


The picture depicts the dendrogram of the hierarchical clustering

```
[173]  df
       ✓ 0.0s                                                    Python
```

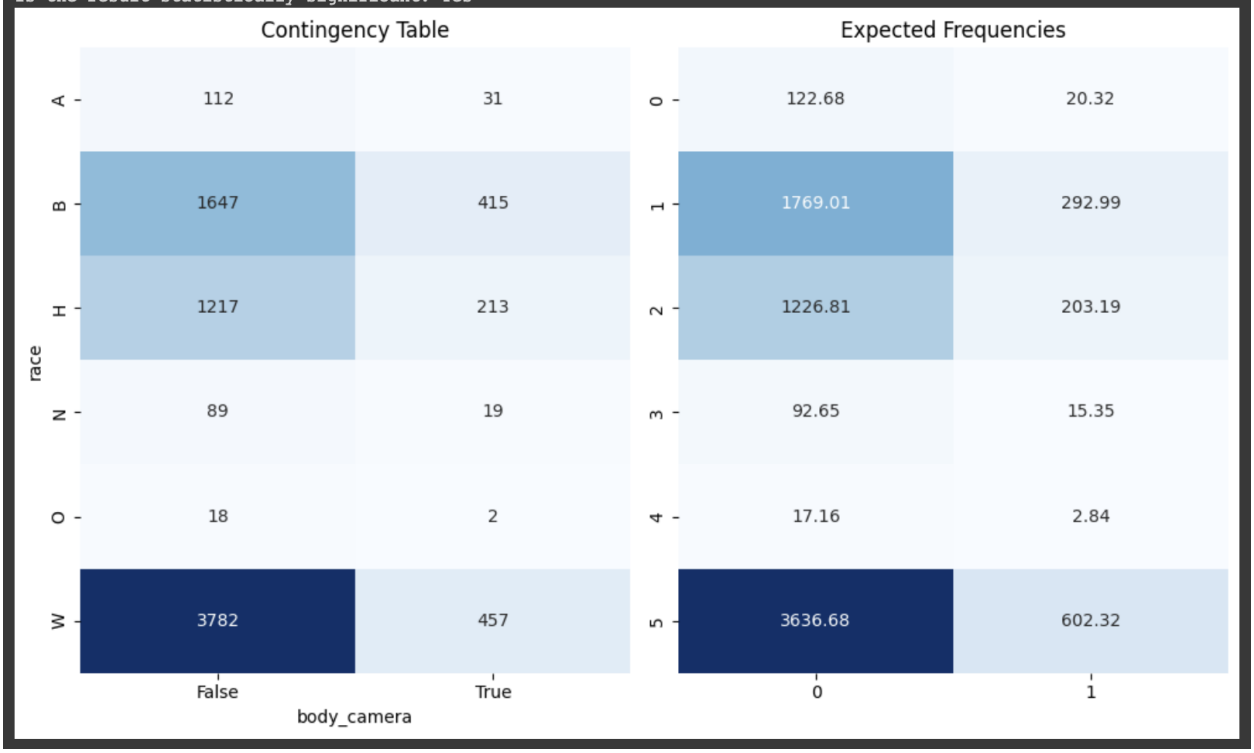
	PredictedClusterID	flee	armed	threat_level	age_group	Count
0	0	car	knife	other	Adult	19
1	1	foot	gun	attack	Adolescents	24
2	2	Notfleeing	unarmed	attack	Adult	10
3	3	foot	gun	other	Adult	13

In this picture: *we can see the trend in the clusters of why the people in these clusters were being shot*



Visualizing the trend of people being shot using the bar graphs

```
Chi-squared statistic: 108.49418618642981
P-value: 8.524478119592931e-22
Degrees of Freedom: 5
Expected Frequencies:
[[1.22681205e+02 2.03187953e+01]
 [1.76901150e+03 2.92988503e+02]
 [1.22681205e+03 2.03187953e+02]
 [9.26543364e+01 1.53456636e+01]
 [1.71582104e+01 2.84178955e+00]
 [3.63668270e+03 6.02317296e+02]]
Is the result statistically significant? Yes
```



The extremely low p-value suggests that there is a significant association between race and body camera status.

Appendix C: Code

The code can be viewed in this GitHub repo
Link: https://github.com/nikhil9066/Police-shooting-MTH_P2

Code 1: Given details of a person, determine the race of a particular person.

```
import pandas as pd
from sklearn import tree
#choose from different tunable hyperparameters
clf = tree.DecisionTreeClassifier(max_depth=3,criterion='entropy')
 #Printing all the parameters of Decision Trees
print(clf)
#Creating the model for Training Data
DTree=clf.fit(X_train,y_train)
prediction=DTree.predict(X_test)
#Measuring accuracy on Testing Data
from sklearn import metrics
print(metrics.classification_report(y_test, prediction))
print(metrics.confusion_matrix(y_test, prediction))
#Plotting the feature importance for the Top 10 most important columns
feature_importances = pd.Series(DTree.feature_importances_, index=Predictors)
feature_importances.nlargest(10).plot(kind='barh')
#Printing some sample values of prediction
TestingDataResults=pd.DataFrame(data=X_test, columns=Predictors)
TestingDataResults['TargetColumn']=y_test
TestingDataResults['Prediction']=prediction
TestingDataResults.head()
# Plotting the decision tree
plt.figure(figsize=(20, 10))
tree.plot_tree(DTree, filled=True, feature_names=Predictors, class_names=True,
rounded=True)
plt.show()
```

Code 2: Find the index of the row with the maximum count for each cluster

```
import matplotlib.pyplot as plt
import seaborn as sns
cluster_incident_info = df_cal.groupby(['PredictedClusterID', 'flee', 'armed',
'threat_level','age_group']).size().reset_index(name='Count')
cluster_incident_info.sort_values(by='Count', ascending=False)
# Find the index of the row with the maximum count for each cluster
max_count_index =
cluster_incident_info.groupby('PredictedClusterID')['Count'].idxmax()
max_count_rows = cluster_incident_info.loc[max_count_index]
max_count_rows
sns.set(style="whitegrid")
fig, axes = plt.subplots(3, 2, figsize=(15, 15))
# Plotting for 'age_group'
sns.barplot(x='PredictedClusterID', y='age_group', data=cluster_incident_info,
palette="viridis", ax=axes[0, 0])
axes[0, 0].set_title('Mode of Age Group')
 sns.countplot(x='threat_level', data=cluster_incident_info, palette="viridis", ax=axes[0,
1])
axes[0, 1].set_title('Mode of Threat Level')
sns.countplot(x='armed', data=cluster_incident_info, palette="viridis", ax=axes[1, 0])
axes[1, 0].set_title('Mode of Armed Status')
sns.countplot(x='flee', data=cluster_incident_info, palette="viridis", ax=axes[1, 1])
axes[1, 1].set_title('Mode of Flee Status')
sns.barplot(x='PredictedClusterID', y='Count', data=cluster_incident_info,
palette="viridis", ax=axes[2, 0])
axes[2, 0].set_title('Count in Each Predicted Cluster')
fig.delaxes(axes[2, 1])
plt.tight_layout()
plt.show()
```

Code 3: The probability of a Male being shot when the person is from the race

```
#Cleaning the data
df=pd.read_excel("fatal-police-shootings-data_original.xls")
df_cal=df[df['state']=='CA']
# Prior probability of being shot
prior_prob_shot = df_cal[df_cal['manner_of_death'] == 'shot and Tasered'].shape[0] /
df_cal.shape[0]
# Prior probability of being male
prior_prob_male = df_cal[df_cal['gender'] == 'M'].shape[0] / df_cal.shape[0]
# Prior probability of being Black
prior_prob_black = df_cal[df_cal['race'] == 'H'].shape[0] / df_cal.shape[0]
# Likelihood of being shot given that the person is male and Black
likelihood_shot_given_male_and_black = df_cal[(df_cal['gender'] == 'M') & (df_cal['race']
== 'H') & (df_cal['manner_of_death'] == 'shot and Tasered')].shape[0] /
df_cal[(df_cal['gender'] == 'M') & (df_cal['race'] == 'H')].shape[0]
# Apply Bayes' theorem
posterior_prob_male_shot_given_black = (likelihood_shot_given_male_and_black *
prior_prob_shot) / prior_prob_black
print(f"The posterior probability of a Male being shot when the person is from the 'H'
race is approximately: {posterior_prob_male_shot_given_black:.4f}")
```

Code 4: was there any bias when there was a body cam on any race

```
contingency_table = pd.crosstab(df['race'], df['body_camera'])
# Perform the chi-squared test
chi2_stat, p_val, dof, expected = chi2_contingency(contingency_table)
# Determine if the result is statistically significant
alpha = 0.05
is_significant = p_val < alpha
print(f"Chi-squared statistic: {chi2_stat}")
print(f"P-value: {p_val}")
print(f"Degrees of Freedom: {dof}")
print(f"Expected Frequencies:\n{expected}")
print(f"Is the result statistically significant? {'Yes' if is_significant else 'No'}")
# Visualization
plt.figure(figsize=(10, 6))
# Contingency Table Heatmap
plt.subplot(1, 2, 1)
sns.heatmap(contingency_table, annot=True, fmt='d', cmap='Blues', cbar=False)
plt.title('Contingency Table')
# Expected Frequencies Heatmap
plt.subplot(1, 2, 2)
sns.heatmap(expected, annot=True, fmt='.2f', cmap='Blues', cbar=False)
plt.title('Expected Frequencies')
plt.tight_layout()
plt.show()
# Findings
# The extremely low p-value suggests that there is a significant association between
race and body camera status.
```