**Group 8: Final Project**

Amber Converse, Kai Shuen Neo, Iris Sum

School of Information, University of Arizona

INFO 523: Data Mining and Discovery

Dr. Hong Cui

May 6, 2024

**Abstract**

This paper analyze the data of 8,720 police shootings happened in the US. Overall, males victims had a higher tendency of being victimized in fatal police shootings, with most of them being of 'White' race, and having 'Gun' as their weapon of choice. Chi-Squared analysis and Cramer's V were used to test the corelation between race and the weapon used, and resulting as no strong corelation observed. Association itemsets and rules are mined and presented in the paper. Random Forest and Decision Tree were built to predict race. Cluster analysis using the PAM algorithm and the OPTICS algorithm were conducted and compared. Both algorithms results in one big cluster with similar attribute characteristic: majority of the observations were white male victims armed with gun.

## Introduction

Using data from a collection of police shootings in the United States, we performed an exploration of the statistics of shootings using a data mining methods for data pre-processing and analysis. We reached a null hypothesis that there is a correlation between race and the weapon used. While a Chi-Squared analysis showed a correlation between race and the weapon used, the results of Cramer's V on that correlation lead us to accept the null hypothesis, showing that there is no correlation between race and the weapon used.

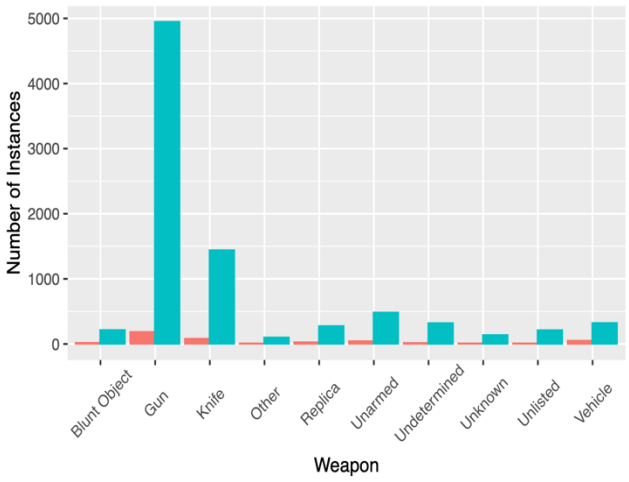
## Data and Methods

The data used was a collection of 8,720 police shootings that happened in the US. Each shooting record had 19 attributes: *id, data, threat_type, flee_status, armed_with, city, county, state, latitude, longitude, location_precision, name, age, gender, race, race_source, was_mental_illness_related, body_camera, agency_ids*. Attributes were organized into 3 data types: numeric, nominal, and binary. Country values were missing in over 50% of cases, but were able to imputed using the US Cities Data CSV to reduce the missing number of values to 7.6%. For our analysis, location was standarized using the state attribute as it reflected regional differences with no data loss. The attributes *gender, threat_type, flee_status, armed_with, age,* and *race* were selected as important attributes for analysis and rows with these attributes missing were removed as imputation was impossible for these attributes.

First, data exploration was performed using summary statistics to define the distribution of the data for selected attributes. Then, data pre-processing was performed by analyzing missing values for attributes and performing a Pearson's correlation coefficient analysis and Chi-Squared correlation test. Next, frequent itemsets were mined using the Apriori algorithm and association rules were generated from those itemsets. Then, a Random Forest model and decision tree were generated to predict race using the selected attributes. Cluster analysis was performed using two methods, PAM and OPTICS. Finally, the data was analyzed using outlier detection to search for univariate, multivariate and contextual outliers.

## Results and Discussions

**Figure 1.** *Police Shootings by Weapon & Gender*




The police shooting dataset showed majority of the victims are male (95.6%), with the mean of both male and female victims lie in between the '41 to 50' age group. *'White' (43%), 'Black'* (23%) and *'Hispanic' (15.1%)* races victims were found to be commonly targeted. Regarding the weapon usage, 5.7% of male victims and 9.6% of female victims are found to be unarmed. 'Gun' and 'Knife' were the most commonly used weapons for both male and female victims (**Figure 1**). For victims with vehicle, 71% of male and 70% of female victims were moving with car.

Correlation and independence between both categorical and continuous variables were then tested to detect relationships between variables and identify redundant attributes in the data set. The Pearson's correlation coefficient analysis on numeric attributes (age, date, latitude, longitude) showed no strong linear correlation between age, date and location coordinates, with P-value < 0.0754 which were likely impacted by the large sample size.

**Table 1.** *Expected Frequency vs. Actual Frequency for Black and White Victims*

| | **Black** | | | **White** | | |
|---|---|---|---|---|---|---|
| **Weapon** | **Expected** | **Actual** | **Difference** | **Expected** | **Actual** | **Difference** |
| **Gun** | 1164.24368 | 1246 | 7.00% | 2212.6006 | 2245 | 1.50% |
| **Melee** | 391.92585 | 306 | -21.90% | 744.8401 | 731 | -1.90% |
| **Weapon Other** | 21.25853 | 22 | 3.50% | 40.40102 | 38 | -5.90% |
| **Replica** | 65.58482 | 55 | -16.10% | 124.64145 | 154 | 23.60% |
| **Unarmed** | 116.24344 | 160 | 37.60% | 220.91622 | 199 | -9.90% |
| **Unknown** | 107.42342 | 78 | -27.40% | 204.1541 | 180 | -11.80% |
| **Vehicle** | 82.32026 | 82 | -0.40% | 156.44651 | 157 | 0.40% |

Correlation between weapon type and race of the victims was found through the Chi-squared analysis (**Table 1**). Black victims were more likely to be unarmed or have a gun, but less likely to have a replica weapon. White victims were more likely to have a replica, but less likely to be unarmed or

have a gun. However, upon testing this correlation with Cramer's V, a value of 0.07 was found, which suggests the correlation is not strong. Thus, the null hypothesis was rejected.

To observe the association between items in the data set, Apriori algorithm was used to mine the 5 longest frequent item sets for male and female victims. Not informative attributes such as id and attributes with high cardinality such as latitude, longitude were removed before mining. The item set with the highest support (0.086) and count (781) for male victims turned out to be: *[gender = male, race = H, race source = not available, was mental illness related = False, body camera = False]*. As for female victims, the item set with the highest support (0.086) and count (36) turned out to be: *[flee status = , gender = female, was mental illness related = False, body camera = False]*.

| **Rules** | **Kulc** | **Imbalance** |
|---|---|---|
| {was_mental_illness_related=False} => {body_camera=False} | 0.8301640 | 0.04830434 |
| {body_camera=False} => {was_mental_illness_related=False} | 0.8301640 | 0.04830434 |
| {threat_type=shoot} => {armed_with=gun} | 0.7316479 | 0.51237978 |
| {armed_with=gun} => {body_camera=False} | 0.7267254 | 0.28118915 |
| {age_group=grown} => {body_camera=False} | 0.7242762 | 0.26017722 |

**Table 2.** *Five Strongest Association Rules by Kulc scores and Imbalance Ratio for male victims*

In order to evaluate the interestingness of these rules. The Kulczynski measure and Imbalance Ratio were evaluated to find the rules with the greatest Kulczynski Measure as shown in Tables 2 and 3. All of these rules for male victims have a Kulczynski Measure of greater than 0.7, suggesting a positive correlation between the left and right hand sides. However, only the rule {threat_type=shoot} => {armed_with=gun} is imbalanced. Only this rule is interesting for male victims out of these 5 rules with the strongest Kulczynski Measure. This follows logically because if the threat type was a shooting, it is likely that the victim of the shooting was the threat; therefore, they were armed with a gun.

| **Rules** | **Kulc** | **Imbalance** |
|---|---|---|
| {was_mental_illness_related=False} => {body_camera=False} | 0.8095356 | 0.1931818 |
| {age_group=grown} => {body_camera=False} | 0.7424397 | 0.2314050 |
| {race=W} => {body_camera=False} | 0.7308580 | 0.3276353 |
| {flee_status=not} => {body_camera=False} | 0.7276302 | 0.2659280 |
| {threat_type=shoot} => {armed_with=gun} | 0.7123397 | 0.5536723 |

**Table 3.** *Five Strongest Association Rules by Kulc scores and Imbalance Ratio for female victims*

For female victims, all of the rules in Table 3 have Kulczynski Measure of greater than 0.7, which also suggests a positive correlation. However, the Imbalance Ratios suggest the only two interesting rules in this group are {race=W} => {body_camera=False} and {threat_type=shoot} => {armed_with=gun}. The first rule that suggests a victim being white leads to a lack of body camera footage being available still has an Imbalance Ratio of close to 0.3, so while it is not balanced, it is

not particularly imbalanced either. The most interesting rule is the same as for male victims: {threat_type=shoot} => {armed_with=gun}.

| Race | Precision | Recall |
|---|---|---|
| **White** | 0.59 | 0.78 |
| **Black** | 0.54 | 0.42 |
| **Hispanic** | 0.45 | 0.3 |
| **Asian** | 0 | 0 |
| **Native American** | NaN | 0 |
| **Other** | NaN | 0 |

**Table 4.** *Random Forest evaluations*

| Race | Precision | Recall |
|---|---|---|
| **White** | 0.57 | 0.89 |
| **Black** | 0.44 | 0.3 |
| **Asian** | NaN | 0 |
| **Hispanic** | NaN | 0 |
| **Native American** | NaN | 0 |
| **Other** | NaN | 0 |

**Table 5.** *Decision Tree evaluations*

A Random Forest model was constructed with data split in train (80%) and test (20%) sets to predict race. No class weights were included as they had negligible effects on performance. Results showed an overall accuracy of 56%, with precision and recall evaluation shown in **Table 4**. The important attributes were evaluated using the attribute importance plot for random forest model. *Age* and *region* were found to be the most important attributes, whereas *gender* and *month* being the least important attributes. This conceptually indicated that gender did not actually predict race and it is unlikely that shootings in different months affected different races differently. This model further support on rejecting the null hypothesis as the attribute *armed_with* was not significantly important for predicting race. A Decision Tree was trained to predict races using only the age attribute and its precision and recall evaluation are shown in **Table 5**. Since there are no predicted result in some races, those precision cannot be computed and result as "NaN". As such, each node in the Decision Tree was either 'White' or 'Black'. An interesting founding from the Decision Tree model was that 100% of race 'Black' had ages under 26.

**Figure 3.** *Clusters result with PAM*

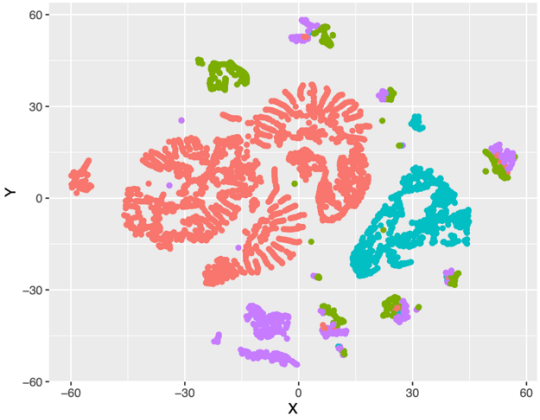



**Figure 4.** *Clusters result with OPTICS*

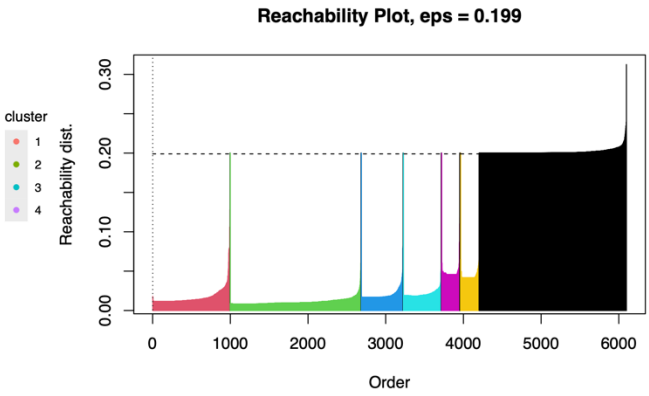



**Figure 3** shows the result of 4 clusters using the PAM algorithm with a K-means of 4, and **Figure 4** shows the result of 6 clusters using the OPTICS algorithm with eps of 0.199. Both algorithms results in one big cluster with similar attribute characteristic: majority of the observations

were white male victims armed with gun. OPTICS algorithm generated clusters with more pure observations' class in each attribute whereas PAM algorithm's clusters contained a mixture of values. For example, in OPTICS clustering results, all 6 clusters were male and there was only 1 race in each cluster. However, in PAM clustering results, the largest cluster was dominated with victims armed with gun, together with a few victims armed with other weapons.

For outlier analysis, global, contextual and collective outliers were identified. For univariate contextual outliers, looking into age attributes, the 79 circle points lie beyond the whiskers may considered as outliers (**Figure 5**). For observations with value different from the typical defined options in the data set may be identified as an global outlier, for example, value "B;H" outside typical defined options "A","W","H","B","O","N" in race attributes.

For contextual outliers, the dataset was divided into 50 subsets with the contextual attribute state to observe if any behavior attribute race can be identified as contextual outlier. Several examples were observe such as race *"Native American"* (1 observation) in state *"TX"* (665 observations) and

**Figure 5.** *Boxplot of Age*

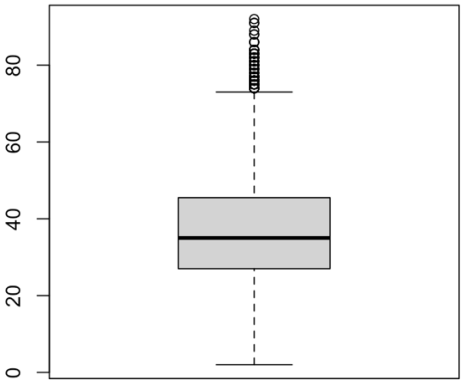
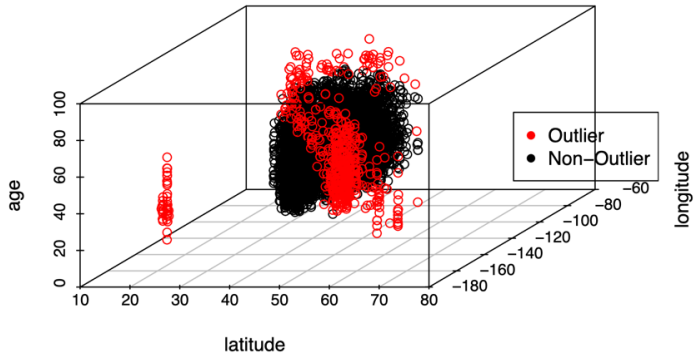



**Figure 6.** *3-D Scatterplots with age, latitude and longitude*




race *"Asian"* (2 observations) in in state *"FL"* (490 observations), knowing that race *"Native American"* or *"Asian"* themselves are not considered as global outliers in the data set.

For collective outliers, an example would be a cluster of victims that as a collective had a mean of 70-years-old with low variance between them. This differs significantly from the overall distribution of age across the country.

To identify the multivariate global outliers, Mahalanobis distance was calculated for age, latitude and longitude, followed by the Chi-square quantile to determine which data points were outliers. 565 outliers in red circle were identified in **Figure 6**. Notably, there were data points around 20N 150W corresponding to the location of Hawaii. These points should not be considered outliers solely because of their position since there was a consistent mechanism that predict why this cluster exists.

## Tentative Conclusion and Future Work

The mining of the police shooting dataset rejected the null hypothesis of having correlation between race and the weapon used. Through the cluster analysis, a majority of white male victims armed with gun were observed. The next step in evaluating this would be to set a new hypothesis to

be test. Is there any correlation exist between other attributes? How the accuracy of the prediction model can be improved? Can the mined association rules help in modifying the prediction model?

Improvement has been made as per the assignment feedback, including but not limited to updating association rules mined with Kulc and imbalance instead of support and confidence, and the contextual outliers identification.

## References

United States Cities Database. Simplemaps. Accessed 5 May 2024. https://simplemaps.com/data/us-cities