**THE TOLL OF POLICE SHOOTINGS IN THE UNITED STATES**

**Co-authors:**

Neeresh Kumar Perla

Pooja Chigurupati

Rahul Manjunath

Nikita Michelle Rego

**The Issues:**

1. To what extent are there racial disparities in fatal police shootings in the United States from 2015 to 2022 using Years of Life Lost (YLL).
2. How do the identified patterns between race, signs of mental illness, and age within the clusters contribute to the overall understanding of incidents in each group?
3. The Significance of Body Camera Presence in Police Shootings
4. To what extent does mental illness play a role in fatal police shootings in the United States

**Findings:**

* **Racial Disparities:**
  + Most of the victims in fatal police shootings during this period were White, accounting for approximately 50.88% of the total. Black individuals comprised 27.03%, followed by Hispanic victims at 18.25%. Asian victims constituted 1.98%, while individuals categorized as "Other" represented 1.485%
  + The median age of victims also exhibited variations across racial groups. Whites had a median age of 35, whereas Black victims were younger, with a median age of 31. Hispanic victims fell in between with a median age of 33, while Asian victims shared a median age of 35. Those categorized as "Other" had a median age of 32.
  + The death of unarmed victims is drastically reduced over the years from 2015 to 2022.
  + The mean years of life of is more for Asians, while blacks and whites is very similar.
* **Cluster Analysis on different clustered groups by race, signs of mental illness, and age:**
  + An analysis to determine the best way to group similar incidents based on data. Using the elbow method, found that the optimal number of groups is 4, which created using the K-means algorithm.
  + The comparison revealed that each group is distinct, with similar incidents in each group. This was supported by a Silhouette score of 0.59. We also determined the central points for each group to understand the characteristics of incidents in each cluster.
  + Found patterns within each group, including a correlation between being armed and perceived threat level, signs of mental illness and tendency to flee, and differences in age distribution across racial groups.
* **Body Camera Analysis:**
  + The analyses affirm the pivotal role of body cameras, in impacting outcomes and perceptions of police shootings.
  + Gender bias without body cameras is a significant concern requiring targeted intervention and reform.
  + There are a lot of cases where the information is unknown or missing when the body camera is not present.
* **Mental Illness:**
* The number of victims of fatal police shootings who showed signs of mental illness decreased by 53.07% from 2015 to 2022
* Between 2015 – 2022, California exhibits the largest count of victims who had mental illness, i.e. 13% of the total number of victims with mental illness; while Rhode Island shows the least number i.e. 0.05%
* In terms of race, the average of white population having mental illness is 64.27% of the number of shootings with mental illness victims, with black victims having an average of 18.28% followed by Hispanic victims of 13.74%
* The victims who showed mental illness, showed a preference for using guns (49.67%) followed by knives (23.09%)
* 77% of the time, victims with mental illness do not flee the scene.

**Discussion:**

* **Racial Disparities:**
  + The relatively high percentage of Black victims (27.03%) suggests a disproportionate representation of Black individuals in fatal encounters with law enforcement. This finding emphasizes the need for further investigation into the root causes and potential reforms in policing strategies and policies.
  + The significant reduction in unarmed victim deaths over the years is a positive finding. It suggests that changes in police training, de-escalation tactics, or policy reforms may be contributing to improved outcomes in terms of minimizing the use of lethal force in situations involving unarmed individuals.
  + The difference in mean years of life lost among racial groups, with Asians showing a higher life expectancy, is a critical observation. This highlights disparities in the potential years of life lost to fatal police shootings and underscores the need for addressing these disparities.

**• Cluster Analysis on different clustered groups by race, signs of mental illness, and age:**

* The identification of four distinct incident clusters provides a clear categorization, allowing law enforcement to tailor responses based on the specific characteristics of each cluster.
* Correlation patterns within Cluster 2 highlight key associations, such as the link between being armed and perceived threat level. This knowledge aids in understanding incident dynamics and potential risk factors.
* The observed correlation between signs of mental illness and the tendency to flee within Cluster 2 underscores the importance of specialized intervention strategies. Law enforcement can focus on de-escalation tactics and mental health support to reduce the likelihood of incidents escalating in this particular cluster.
* **Body Camera Analysis:**
  + The observed significant decrease in police shootings after 2022 implies positive changes, potentially attributed to heightened awareness, policy advancements, or shifts in law enforcement practices. This decline raises optimism about evolving strategies aimed at reducing the frequency of such incidents.
  + The results of chi-squared tests underscore the pivotal role of body cameras in shaping various aspects of police encounters. These findings emphasize the significance of employing body cameras as a tool for enhancing transparency, accountability, and overall policing practices.
  + The identified gender bias in cases without body cameras highlights a critical issue that demands immediate attention and reform within law enforcement practices. The presence of such bias underscores the importance of implementing policies and training that promote gender-neutral and fair treatment in police interactions.
* **Mental Illness:**
  + The decrease in number of victims who displayed signs of mental illness suggests that there are certain measures and protocols in place to deal with such people to lessen the victim count.
  + California in general has a high body count. But the fact that a vast number of those victims showed signs of mental illness would imply that not enough measures are put in place while confronting such people.
  + The fact that a larger portion victim pool is white population having mental illness, along with the facts that the white population tend not to flee the scene and people of mental illness tending to not flee, would indicate that a cop has a higher chance of encountering a person with mental illness if they are White and do not flee.

**Appendix A: Method**

In examining racial disparities in 5194 fatal police shootings from 2015 to May 2022, we observed a decline in unarmed fatalities, particularly among Black individuals, with White victims comprising 50.88%. Generalized Linear Models (GLMs) explored relationships, revealing insights such as the impact of age, latitude, gender, and race. Cluster analysis, employing K-means clustering, identified four clusters based on race, signs of mental illness, and age, showcasing well-separated groups with a silhouette score of 0.59. Analysis of 8002 cases regarding body cameras exposed significant associations with race, gender, armed status, and fleeing behavior. Logistic regression achieved 84.61% accuracy in predicting cases without body cameras. Investigating 1671 police shootings involving mental illness, we unveiled trends, demographics, and state-wise distributions. Logistic regression and Random Forest models predicted mental illness with 78% accuracy post-hyperparameter tuning. The holistic approach provides comprehensive insights into police shootings, emphasizing the need for transparency, understanding disparities, and predictive modeling for mental health indications.

**Appendix B: Results**

1. **Racial Disparities:**

There were 8002 fatal police shootings reported by the Washington Post from 2015 to May 2022. A total of 2808 was excluded from the analysis because race was unknown or belong to Native Americans or marked as ‘other’ and the age was not reported for 135 victims, leaving 5194 deaths included in the YLL analysis. Most victims were White (50.88%), followed by Black (27.03%), Hispanic (18.25%), Asian (1.98%) and Others (1.485%). Fig1.1 shows the visualization of the percentage of victims that are killed across each racial group. The median age of victims was 35 and varied across groups with White victims (35), Black victims (31), Hispanic victims (33), and Asian (35).

A graph with different colored squares

Description automatically generated

Fig1.1: Percentage of victims that are killed across each racial group.

The decrease in the loss of life among unarmed victims is remarkable. Figure 1.2 illustrates the trends for each racial group from 2015 to 2022. In 2015, unarmed Black individuals had the highest number of fatalities, followed by White individuals. However, as the years have passed, a substantial decline has become evident, with the rate dropping to single digits. This decline suggests that significant steps have been taken by law enforcement to address this issue.

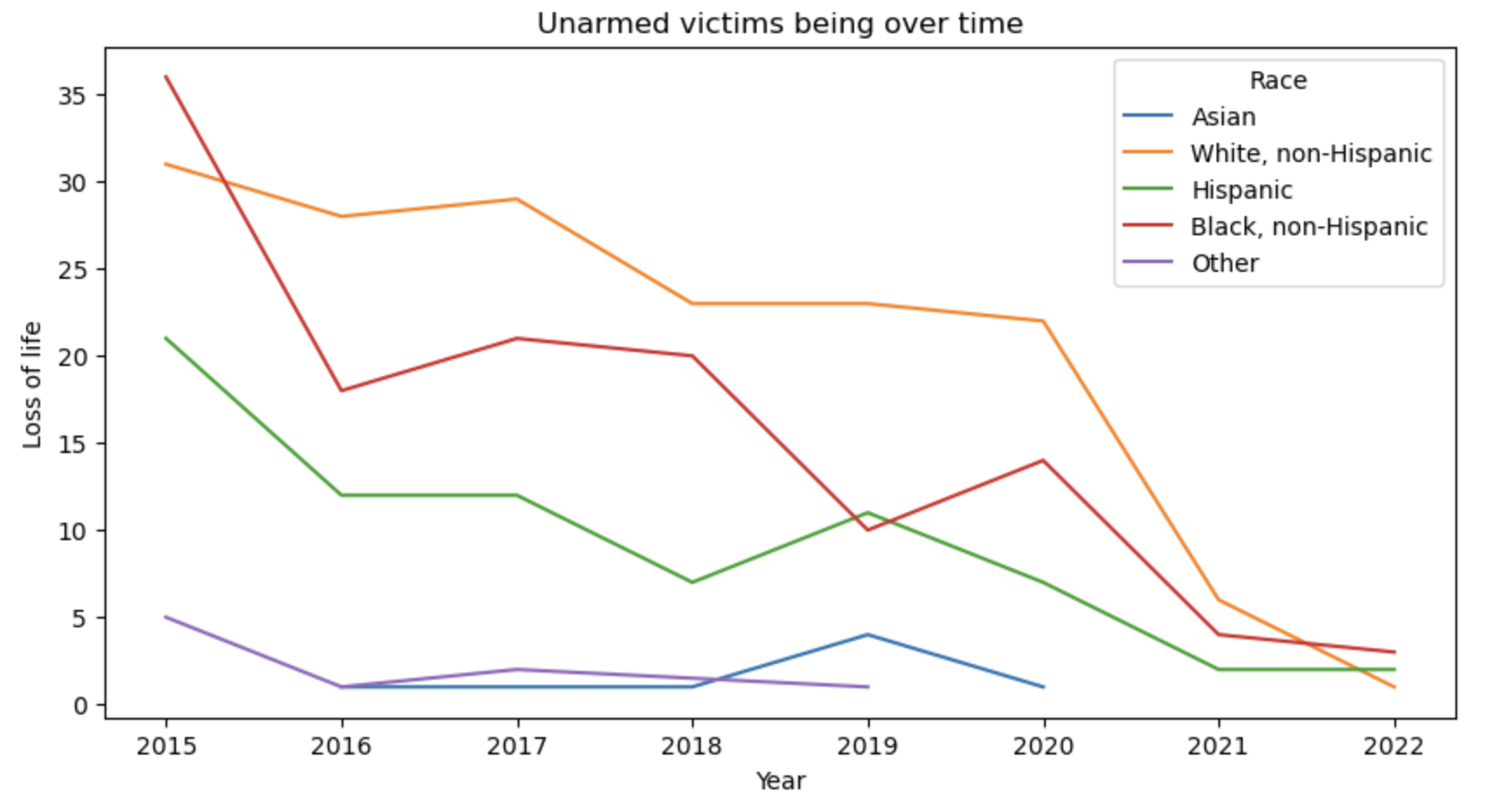


Fig 1.2: Kill trend of unarmed victims from 2015 to 2022

The average years of life lost are nearly the same for the Asian-Hispanic and White-Black racial groups. The trend lines for the White-Black racial groups have converged closely from 2015 to 2021 and have significantly decreased at the beginning of 2022. In contrast, the average years of life lost are notably higher for the Asian racial group, with the highest peaks recorded in 2017 and 2021. There is a significant decrease in the average years of life lost, reaching its lowest point in 2021 when compared to the other groups, and it has slightly increased at the beginning of 2022. Figure 1.3 illustrates this:

**A graph of different colored lines

Description automatically generated**

Fig 1.3: Mean trend of years of life lost across racial groups from 2015 to 2022

Figure 1.4 clearly exhibits inconsistencies over the years. The variation in the other racial group increased significantly from 2017 to 2020 and then experienced a drastic drop in 2022. In contrast, there are minimal fluctuations in the years of life lost for White, non-Hispanic individuals, and this figure is consistently decreasing. Additionally, there is a recurring pattern of variance in the Asian racial group from 2015 to 2021.

A graph with different colored lines

Description automatically generated

Fig 1.4: Variance of YLL across racial groups from 2015 to 2022

All the victims were armed during the shooting, and a total of 95 distinct items carried by the victims have been categorized into 8 groups, as shown in the Fig 1.5, to perform deeper analysis.

A white background with red text

Description automatically generated

Fig 1.5 Categorization of 95 different weapons into 8 categories

GLMs are extension of linear regression and are commonly used for regression and classification tasks. They main idea behind GLMs is to model the relationship between a response variable and one or more predictor variables while allowing for different pdfs and link functions. Link function is used to connect the linear predictor to the mean of the repones variable. It can be chosen based on the nature of the data and the relationships that we want to model. For this data, we chose "Identity". IRLS (Iteratively Reweighted Least squares) is the optimization method I used to estimate the coefficients. The summarized results of the GLMs is given in the figure 1.6.

A screenshot of a computer

Description automatically generated

Fig 1.6: Generalized Linear Model Results

**Interpretations:**

1. **Intercept (0.5468):** This is the expected value of the response variable "yll" when all predictor variables are zero. In this context, it represents the baseline or starting value of "yll" when all other factors are absent or equal to zero.
2. **age (-1.0039):** For each one-unit increase in the "age" of the individual, the expected value of "yll" is expected to decrease by approximately 1.0039 units. This suggests that older individuals tend to have lower values of "yll."
3. **latitude (0.0007):** A one-unit increase in "latitude" is associated with an increase of 0.0007 units in the expected value of "yll." This means that moving north (increasing latitude) is associated with a slight increase in "yll."
4. **gender\_M (-0.4839):** If the gender of the individual is male ("gender\_M" is 1), the expected value of "yll" is expected to be approximately 0.4839 units lower compared to a female ("gender\_M" is 0), holding all other variables constant.
5. **race\_A (0.6247):** Individuals of race "A" are expected to have a "yll" value that is approximately 0.6247 units higher than individuals of other races (e.g., compared to "race\_W," "race\_H," and "race\_B"), holding all other variables constant.
6. **body\_camera\_False (0.0010):** If "body\_camera\_False" is true (1), it's associated with an expected increase of 0.0010 units in "yll" compared to when "body\_camera\_False" is false (0), holding all other variables constant. This suggests that when body cameras are not used, "yll" is slightly higher.
7. **Degrees of Freedom Residual (df residual = 5169):** It represents the number of data points that can vary freely once our model's parameters have been estimated. For this model 5169 degrees of freedom are left after fitting the model, which suggests that the model is relatively simple in terms of parameter estimation, given the large dataset size.
8. **Degrees of Freedom Model (df model = 24):** It indicates the number of parameters or predictor variables in your GLM model. We have 24 predictor variables included in the model.
9. **Log-Likelihood (13462.0):** The log-likelihood value is a measure of how well the GLM fits the data. A higher log-likelihood value suggests a better fit, and the value of 13462.0 indicates a relatively good fit to the data.
10. **Deviance (1.7):** The deviance is a measure of the goodness of fit, with a smaller value indicating a better fit. For deviance 1.7 suggests that the model provides a reasonable explanation of the data.
11. **Pearson Chi-squared (1.71):** The Pearson chi-squared statistic is another measure of goodness of fit. It's similar to the deviance and provides additional information on how well the model fits the data. A value close to the deviance suggests a reasonable fit.
12. **Pseudo R-squared (1.00):** It suggests that our model explains all the variance in the data.

A strong and positive linear pattern is evident when comparing the original values with the model's predictions, signifying a clear and favorable linear relationship between the two. This pattern suggests that the model's forecasts are in close alignment with the actual data, indicating that the model is performing well without any systemic bias. Furthermore, the residuals, or the differences between the observed values and the model's predictions, appear to be randomly scattered around zero. This randomness in the distribution of residuals underscores the model's ability to capture and explain the variation in the data effectively.

**A graph of a graph

Description automatically generated with medium confidence**

Fig 1.7 Scatterplot between True YLL and Predicted YLL

1. **Cluster Analysis on different clustered groups by race, signs of mental illness, and age:**

* Applied the elbow method to identify the optimal number of clusters by plotting the within-cluster sum of squares (inertia) for different values of k. ‘4’ the point where the rate of decrease sharply changes, forming an "elbow" in the plot.

**A graph with a blue line

Description automatically generated**

Fig 2.1 Elbow Method Plot

* By elbow method the optimal number of clusters for incident data is 4. Applied the K-means algorithm with k=4 means on entire feature in dataset that the algorithm will identify and assign each data point to one of these four clusters. Created a scatterplot where each point represents a data point and points are colored according to their assigned cluster.

**A graph of a number of clusters

Description automatically generated**

Fig 2.2 k-means Clusters

* Minimize the sum of squared distances between data points and their assigned centroids.

Centroid 1: [-37.49662983 17.78552884]

Centroid 2: [10.91905054 14.92544861]

Centroid 3: [ 8.71826487 -13.61928805]

Centroid 4: [-39.67903378 -11.18956059]

A graph showing a number of clusters

Description automatically generated

Fig 2.2 k-means Clusters Centroids

* A Silhouette score of 0.59 suggests that the clusters are well-separated and the objects within the clusters are relatively homogeneous.

A screenshot of a graph

Description automatically generatedA graph of a bar chart

Description automatically generated with medium confidence

Fig 2.4 Distance between Clusters and Datapoint Distributions

* A heatmap of the intercluster distances between different clusters. Each cell in the heatmap represents the distance between the corresponding clusters. The pairwise distances between the centroids of different clusters.
* The bar plot provides a quick overview of the distribution of data points among different clusters.

A screenshot of a graph

Description automatically generatedA graph of different sizes and colors

Description automatically generated

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

A graph of a cluster of data

Description automatically generated with medium confidenceA diagram of a body camera distribution

Description automatically generated

Fig 2.5 Distribution of feature in each Cluster

* This histograms plots, each showing how a specific feature's values are distributed within each cluster. It helps visualize whether there are differences in the distribution of that feature across different clusters.
* Histograms for a particular feature in your dataset, looking at each unique cluster separately. For each cluster, a histogram is made to show how the values of a specific feature are spread out within that cluster.

A graph of a box plot

Description automatically generated A graph of a cluster

Description automatically generated A graph of a cluster

Description automatically generated with medium confidence

A graph of a body plot

Description automatically generated

Fig 2.6 Box Plot for each feature in each Cluster

* Boxplots for each feature and categorizing them by clusters, can visually compare the distribution of each feature across different clusters. The boxplots provide insights into the central tendency, spread, and potential outliers within each cluster for each feature.
* 'race', 'threat\_level', 'state', and 'flee' are determined that there are no outliers, it implies that the values within these features in each cluster are relatively consistent, and there are no extreme values that fall significantly outside the typical range.

A close-up of numbers

Description automatically generated

Fig 2.7 statistical summaries

* statistical summaries of each feature across different clusters, giving insights into the distribution and variation of each feature within each cluster.

A group of graphs with different colored bars

Description automatically generated with medium confidence A group of different colored bars

Description automatically generated A group of different colored bars

Description automatically generated A group of graphs showing different sizes and colors

Description automatically generated with medium confidence

Fig 2.8 Countplot for each feature in each cluster

* Countplots for each categorical column in the dataset. Each countplot provides a visual representation of the distribution of features within a specific column, helping us understand the frequency of different values in variables.

A screenshot of a computer screen

Description automatically generated

Fig 2.9 Correlation matrix for cluster2

* Identified the cluster with the highest average correlation, and then find the correlation matrix for that specific cluster using a heatmap.
* Armed and Threat Level (0.216): There is a moderate positive correlation between being armed and the perceived threat level. On average, individuals who are armed may be perceived as posing a higher threat within this cluster.
* Signs of Mental Illness and Flee (0.160): There is a moderate positive correlation between signs of mental illness and the tendency to flee. On average, individuals with signs of mental illness may be more likely to attempt to flee the scene within this cluster.
* Race and Signs of Mental Illness (0.160): There is a moderate positive correlation between race and signs of mental illness. On average, there may be differences in the prevalence of signs of mental illness across different racial groups within this cluster.
* Race and Age (0.273): There is a moderate positive correlation between race and age. On average, there may be differences in the age distribution across different racial groups within this cluster.

1. **Body Camera:**

* A significant decline in police shootings post-2022, suggests positive shifts in law enforcement practices.

A graph with a line and a blue line

Description automatically generated

Fig 3.1 Yearly trend of the number of cases

* The Chi-squared test results for the relationship between body camera and each of these parameters-race, gender, armed and flee. In the case of race, flee and armed the relationship is highly significant. The relationship between body camera and gender is moderately significant.

A screenshot of a computer screen

Description automatically generated A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated A screenshot of a computer screen

Description automatically generated

Fig 3.2 Relationship of presence of body camera with race, gender, armed and flee

* The hypothesis testing reveals compelling evidence of gender bias when body cameras are absent, supported by a notably high Z-statistic of 994.78 and a low p-value. Despite the statistical significance, the effect size suggests a limited practical impact. Notably, females experience a bias of 6726 more cases than males in incidents where body cameras are not present, emphasizing the gender disparity in the absence of body camera.

A black background with white text

Description automatically generatedA black background with white text

Description automatically generated

Fig 3.3 Z-test of gender bias hypothesis

* The hypothesis testing indicates a substantial and statistically significant difference in 'Unknown' values when a body camera is present, supported by a high Z-statistic of 39.70 and a low p-value. This underscores the significant impact of body cameras on data transparency, suggesting that the presence of body camera contributes to a more complete and comprehensible record of incidents.

A black background with white text

Description automatically generated

Fig 3.4 Z-test of Unknown values hypothesis

* The logistic regression model demonstrates an overall accuracy of 84.61%, indicating its effectiveness in predicting cases without body cameras.

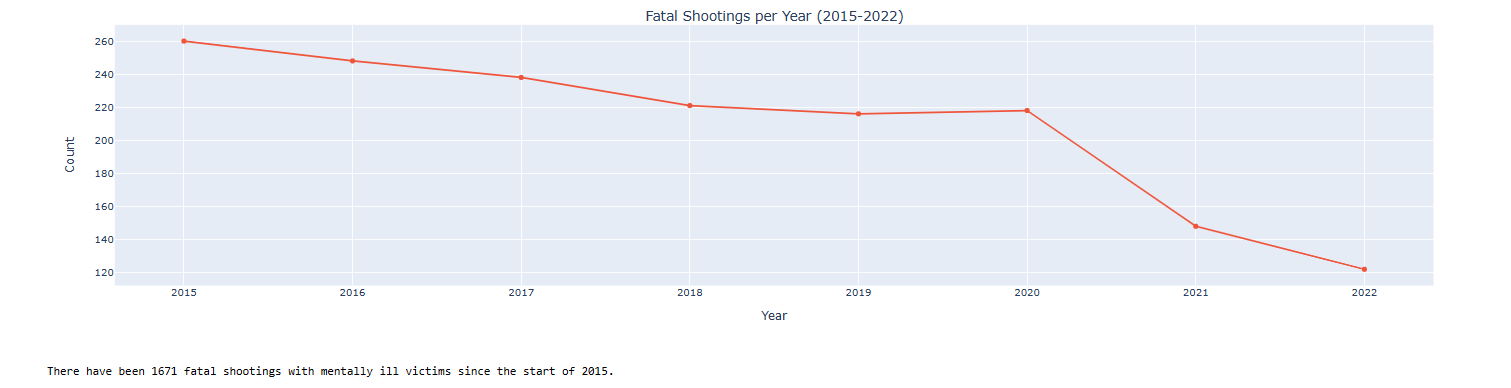
**A screenshot of a computer

Description automatically generated**

Fig 3.5Summary of the Logistic Regression model

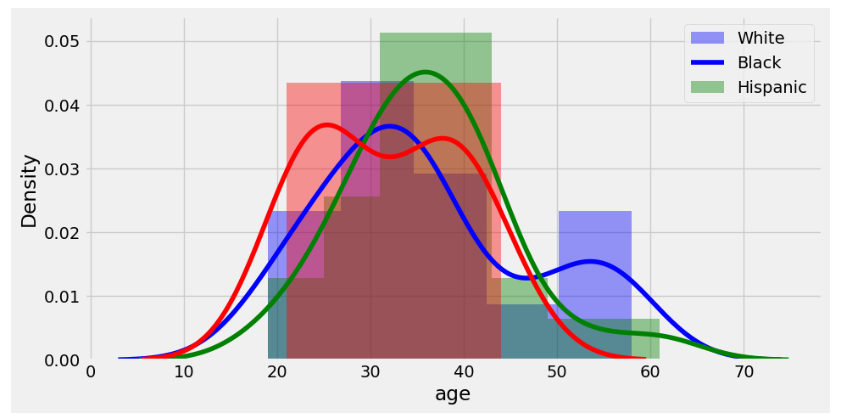
1. **Mental Illness:**

* There were 1671 fatal police shootings involving victims who had mental illness reported by the Washington Post from 2015 to May 2022.

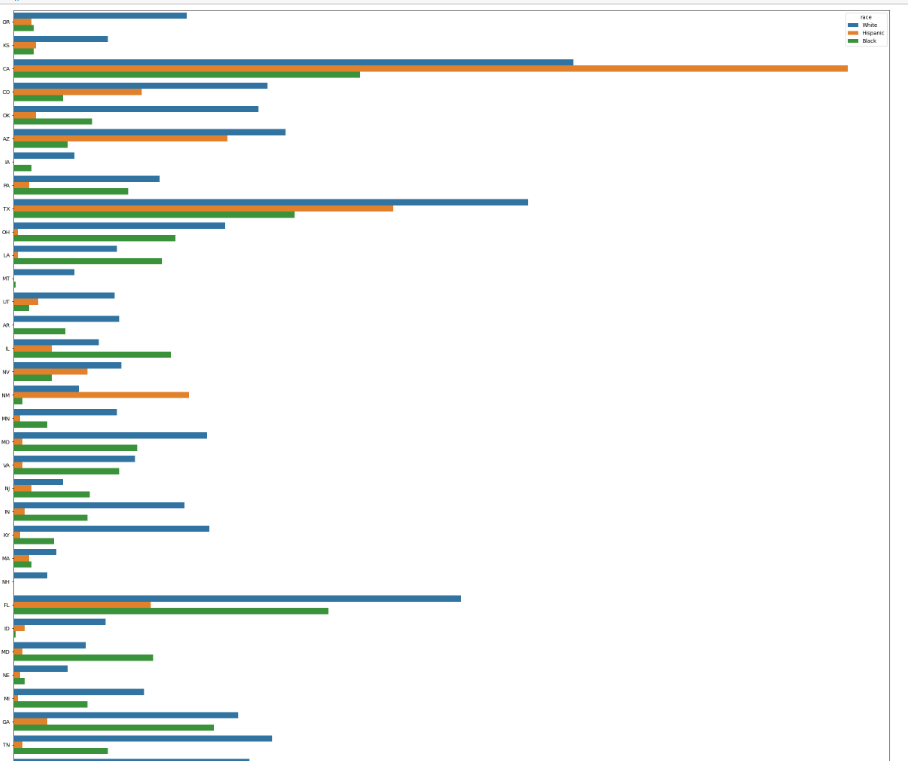


4.1 Trend of shootings involving mentally ill victims

* When we check density of the ages of unarmed victims who displayed signs of mental illness by race, we see that for white population it is left skewed with maximum ages being between 25-60. For blacks the range is 25 to 45 and for hispanics it is 20 to 40.

  
Fig 1.2 Density plot of ages of mentally ill victims w.r.t race

* When we check the victim count by state for races who displayed signs of mental illness, California stands out with over 110 victims who were Hispanic followed by white victims and then Black victims

  
Fig 1.3 Distribution of mental illness victims by race and state

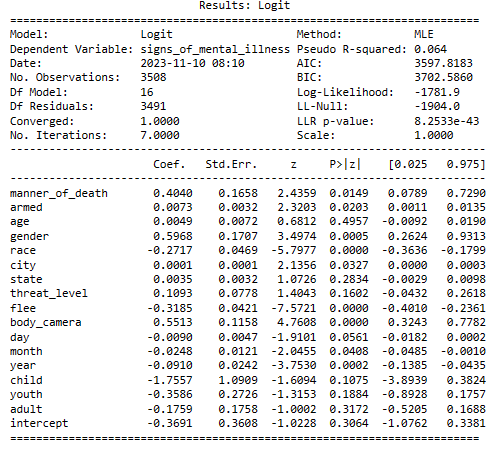
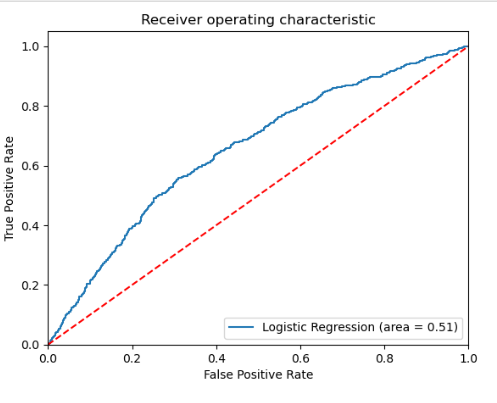
* With the ultimate aim to predict mental illness in the victims, we initiated a logistic regression model and plotted the ROC curve to obtain the following results  
    
   

Fig 1.4 Logistic regression w.r.t mental illness Fig 1.5 ROC curve n w.r.t mental illness

* In creating a Random Forest model to predict illness, we got a model with and accuracy of 77%

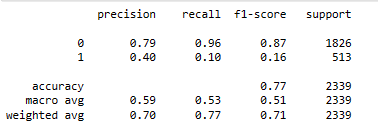


Fig 1.6 Random Forest model

* To improve the model, we tuned the Hyper Parameters and the improved model had an accuracy of 78%

**Appendix C: Code**

The following code is done using Python. **Note:** This is the not the entire code of our project. Each member of our group uploaded their code to their respective GitHub. Links will be mentioned at the end of this report.