Advanced Mathematical Statistics MTH - 522
Project 2

Analysis of Fatal Police Shootings: Clustering and Insights

Authors

Bhanu Prasad Thota
Naga Venkata Lokeswarao Maturi
Mantena Harsha Vardhan Varma
Lakkamraju Hitesh Kashyap Varma

1. The Issues:

The project investigates various aspects of fatal police shootings in the United States, employing demographic analysis and clustering techniques to derive insights. The following questions address key findings and observations in the areas of demographic analysis and clustering.

- Which racial group exhibits the highest number of fatal police shootings?
- How does the graph depict the distribution of deaths based on military or civilian status?
- What is the most prevalent cause of death in police encounters, according to the graph?
- In what age range do the majority of police shooting victims fall?
- Which gender outnumbers the other across all racial groupings in fatal police shootings?
- How does the graph indicate a higher death rate among individuals exhibiting false indicators of mental illness?
- What does the graph depict regarding racial mortality in the United States based on threat level?
- Which mode of transportation is most commonly involved in deadly accidents, according to the graph?
- What percentage of deaths is associated with events having body cameras, according to the research?
- How does the elbow method guide the determination of the optimal number of groups in data organization?
- What do silhouette scores indicate about the effectiveness of various clustering methods in forming well-defined groups?
- How do the KMeans, Hierarchical, and K-Medoids methods visually depict clusters with varying numbers of nodes?
- What intriguing tendencies are revealed in the cluster distribution by race in fatal police shootings?

2. Findings:

a. Demographic Analysis:

The demographic analysis offers a comprehensive understanding of the distribution of fatalities across various demographic factors. The bar plots provide visual insights into the following key aspects:

Race:

The distribution of fatal police shootings across different racial groups is depicted, allowing for an examination of potential disparities. It shows that more white people die in the United States than people of any other race.

Armed Status:
Insights into whether people in situations were armed or unarmed provide important context for understanding the nature of these encounters, particularly in regard to civilian deaths in armed conflicts. This information raises awareness of a developing problem and proposes alternative solutions, such as peaceful conflict resolution, support for disarmament, civilian protection, and accountability for war criminals. The graph depicts deaths based on military or civilian status, indicating that civilian casualties outnumber military casualties. Terrorists and other non-state armed groups are contributing to the escalating trend.

Manner of Death:
The graph depicts how people died in police encounters, the majority of which were caused by gunshots. Gunshot wounds are the most prevalent, followed by multiple gunshot wounds, whereas Taser-related deaths are less common. This brings to light a serious problem in police shootings. Better officer training in de-escalation, increasing responsibility for excessive force, and investing in affected communities to address core problems such as poverty and inequality are among the solutions.

Age:
The graph depicts the age distribution of victims killed by police in the United States. Most deaths occur between the ages of 20 and 30, with gunshot wounds being the leading cause. The research underlines the importance of improved police de-escalation training, increased accountability, and investment in affected communities. The findings highlight the need of addressing police shooting issues and campaigning for policies to decrease such incidences.

Gender:
According to data on fatal police shootings, males outnumber females across all racial groupings, with white males having the highest numbers. Risky behavior and dangerous jobs for men, as well as the impact of structural racism, are potential factors. Public health campaigns, workplace safety measures, and tackling prejudice are among the proposed answers.

Signs of Mental Illness:
The graph shows a greater death rate among people who exhibit false indicators of mental illness, particularly among Black and Hispanic people. It raises the possibility of issues such as violence or preventable factors contributing to these discrepancies. However, particular causes of death are unknown, emphasizing the need for additional study to address health outcome disparities.

Threat Level:
The graph depicts racial mortality in the United States based on threat level ("attack," "other," and "undetermined"). Although it lacks clarity on category meanings and a legend, it shows that attacks are a significant cause of mortality for all races, with relatively fewer deaths classified as "undetermined." Including category data and a legend would improve comprehension.

Flee Status:
The graph depicts the number of fatalities caused by various types of transportation. Cars are the most common mode of transportation involved in deadly accidents, followed by pedestrians and other modes of transportation such as bicycles, motorbikes, and buses. The graph does not represent the hazard level of each mode because it does not consider how frequently each mode is utilized. However, evidence does imply that car accidents kill more people in general, and pedestrians are at a larger risk than those who use other means of transportation.

Body Camera Presence:
According to the research, events with body cams have a lower percentage of deaths (14.6%) than incidents without them (85.4%). Another graph shows a concerning rise in overall mortality in the United States, particularly from COVID-19, with greater rates among Black and Hispanic people due to systematic racism affecting access to healthcare, housing, and education. To solve this, we must enhance healthcare access, engage in marginalized communities, and battle institutional racism to achieve equality of opportunity.

b. Clustering:

The application of KMeans, Hierarchical Clustering, and K-Medoids algorithms aims to identify spatial patterns in fatal police shootings based on geographical coordinates (longitude and latitude). Key findings include:

Elbow Method:
When organizing data, the elbow approach acts as a guide to assist us determine the optimal number of groups. This graph drops quickly until it reaches four groups, at which point it calms down. Look, after four groups, adding more doesn't really make things much better, says the elbow technique. So, four groupings are most likely the sweet spot for our data organization.It's like finding the sweet spot where having more groups won't provide you much more knowledge.

Silhouette Scores:
The silhouette scores allow us to evaluate how well various clustering methods work in forming separate groups in our data. A higher silhouette score indicates that the clusters are better defined. Based on our findings, hierarchical clustering has the maximum score of 0.33, indicating that it generates well-separated groupings. With a score of 0.16, KMeans comes in

second, suggesting decent cluster quality. K-Medoids, on the other hand, trails behind with a lower score of 0.08, signifying fewer unique clusters. In conclusion, based on our data, hierarchical clustering appears to be the most effective in forming well-defined groupings.

c. Cluster Visualization:

The depiction of clusters using the KMeans, Hierarchical, and K-Medoids methods with varied numbers of nodes reveals fascinating patterns. As the number of nodes increases, so does the number and size of clusters, resulting in more complex and irregular structures. KMeans is straightforward but sensitive to beginning conditions, whereas Hierarchical clustering is versatile but computationally demanding. K-Medoids is both durable and computationally demanding. The optimum method is determined by the unique application, considering criteria such as simplicity, computing efficiency, and resilience.

d. Cluster Distribution:

An examination of the cluster distribution by race reveals some intriguing tendencies in fatal police shootings. Two plots show that persons of the same race prefer to cluster together, demonstrating a significant clustering by race. The hierarchical clustering map indicates more equally dispersed clusters dominated by one race, whereas the K-means clustering graphic indicates more racially mixed clusters. Both graphs indicate that factors like as prejudice and personal preferences influence where persons of the same race dwell. A snapshot using the k-medoids clustering technique also shows that the most common group size for all races is 2, with Black and Hispanic populations having more diversified group sizes than White and Asian ones. This means that social structures for Black and Hispanic people will be more diverse. Overall, these findings help to provide a more nuanced understanding of the complicated factors that influence the distribution of fatal police shooting occurrences.

3. Discussion:

a. Demographic Disparities:

An examination of the demographics of fatal police shootings reveals significant differences, particularly in terms of ethnicity, gender, age, and indicators of mental illness. These disparities point to potential policy interventions to address systemic factors that contribute to unequal results. The overrepresentation of specific racial and age groups, for example, highlights the need for focused measures to reform policing methods, improve training in de-escalation strategies, and address structural disparities. Furthermore, the findings concerning armed status, way of death, and threat level highlight the necessity of assessing law enforcement approaches, pushing for improved police training, and investigating options to limit the use of fatal force.

b. Spatial Patterns:

Clustering techniques identify spatial patterns that provide information on the geographic concentrations of fatal police shootings. These trends can be used to guide resource allocation, community policing techniques, and interventions in high-incidence areas. Understanding spatial dynamics is critical for policymakers and law enforcement organizations to devise targeted policies that take into account the differences between locations. The ramifications of geographical patterns extend beyond law enforcement techniques to broader societal issues such as economic inequities, community dynamics, and resource access.

While the study gives light on demographic gaps and spatial patterns, it is critical to acknowledge its limitations. The lack of precise contextual information for each episode, as well as certain critical variables, limits the breadth of comprehension. The findings provide a foundation for future study that should include qualitative data and a broader collection of variables. Such thorough research can contribute to the establishment of equitable and just law enforcement procedures by providing a sophisticated knowledge of the fundamental causes of fatal police shootings.

Finally, the findings highlight the importance of addressing systemic factors that contribute to demographic disparities and spatial patterns in fatal police shootings. Policy reforms, community participation, and targeted interventions are required to improve law enforcement impartiality, transparency, and accountability, eventually building a safer and more equitable society.

c. **Limitations:**

While the analysis provides valuable insights, it is essential to acknowledge its limitations. The dataset, while rich in certain aspects, lacks detailed contextual information for each incident. Critical factors such as the circumstances leading to the encounter, the presence of weapons, and the actions of the individuals involved may influence the outcomes but are not fully captured. Without this context, the analysis can only offer a partial understanding of the complexities surrounding fatal police shootings.

Moreover, the absence of certain variables that could potentially influence the outcomes, such as socioeconomic status, education level, and mental health status, limits the depth of the analysis. Including these variables in future studies could contribute to a more comprehensive understanding of the factors contributing to fatal police shootings.

In conclusion, while the analysis has uncovered demographic disparities and spatial patterns, it serves as a starting point for a more in-depth inquiry. Further research, ideally incorporating qualitative data and a broader set of variables, is crucial for developing targeted and effective strategies to address the root causes of fatal police shootings and promote equity and justice in law enforcement practices.

4. Appendix A: Method

1. Data Collection:

The dataset used in this analysis is obtained from an Excel file that compiles information on fatal police shootings. The Excel file, located at the specified path "D:\MTH\PROJECT2\fatalpolice-shootings-data.xls," serves as the primary source for the investigation. The dataset likely includes various attributes related to each fatal incident, providing a basis for further analysis.

2. Variable Creation:

To facilitate the clustering analysis, certain preprocessing steps are applied to the dataset. Notably, label encoding is employed on categorical variables within the dataset. Categorical variables such as 'manner_of_death', 'armed', 'gender', 'race', 'signs_of_mental_illness', 'threat_level', 'flee', 'body_camera', and potentially others are transformed into numerical representations. This transformation allows for the inclusion of categorical variables in the clustering algorithms, which typically operate on numerical data.

3. Analytic Methods:

a. KMeans Clustering:
The KMeans algorithm is utilized as a spatial clustering technique. The algorithm partitions the dataset into 'k' clusters based on the similarity of data points. The number of clusters, 'k,' is determined through the elbow method, which is visually inspected to identify the point where additional clusters do not significantly reduce within-cluster sum of squares (WCSS).

b. Hierarchical Clustering:
Hierarchical Clustering is employed to explore hierarchical relationships within the dataset. The agglomerative approach is taken, where individual data points are successively combined into clusters. The resulting hierarchy is then used to form a dendrogram, assisting in the determination of an optimal number of clusters.

c. K-Medoids Clustering:
K-Medoids is another clustering algorithm applied for spatial analysis. In contrast to KMeans, KMedoids employs actual data points (medoids) as cluster representatives, making it less sensitive to outliers. The optimal number of clusters is determined similarly to KMeans.

d. Silhouette Scores:
Silhouette scores are calculated to evaluate the goodness of clustering. These scores measure how well-separated clusters are and range from -1 to 1. A higher silhouette score indicates better-defined clusters. The scores are computed for each clustering algorithm (KMeans, Hierarchical Clustering, and K-Medoids), providing a quantitative metric for comparison.

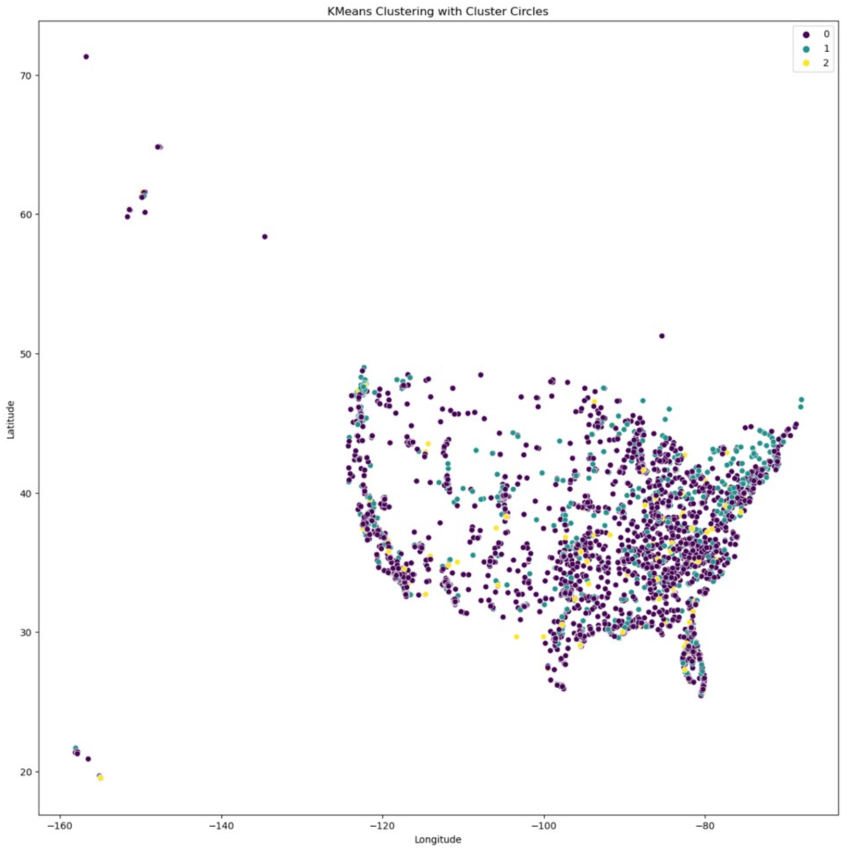
In summary, the methodology involves transforming categorical variables, applying three distinct clustering algorithms for spatial analysis, and assessing the quality of clusters using silhouette scores. These steps collectively contribute to the identification of patterns and structures within the fatal police shootings dataset.

5. Appendix B: Results

a. Clustering Results:

Scatter Plots:
Scatter plots are generated for each clustering algorithm, illustrating the spatial distribution of data points based on longitude and latitude. These plots provide a visual representation of how the algorithms group incidents geographically. The color-coded clusters make it easy to identify distinct patterns and concentrations.

Cluster Circles:
To enhance the understanding of spatial clusters, cluster circles are plotted on the scatter plots. Each circle represents the standard deviation of the geographical coordinates within a cluster. This visualization aids in identifying the spatial extent and density of each cluster, offering insights into the geographical patterns of fatal police shootings.



```
Silhouette Score (KMeans): 0.16134790469330992
```

Distribution Analysis:

The distribution of race within each cluster is analyzed using count plots. This analysis provides an overview of how different racial groups are distributed across the identified clusters. Examining the distribution within clusters allows for the identification of potential disparities or patterns related to race.

b. Silhouette Scores:

Silhouette scores are calculated for each clustering algorithm—KMeans, Hierarchical Clustering, and K-Medoids. These scores measure the cohesion and separation of clusters, providing a quantitative assessment of the quality of the clustering. A higher silhouette score indicates that the data points within a cluster are close to each other and far from points in other clusters.

The silhouette scores are reported to validate the effectiveness of the clustering algorithms in creating distinct and well-separated clusters. This information is crucial for assessing the reliability of the spatial patterns identified through clustering.

```
Silhouette Score (KMeans): 0.1538740220873823
Silhouette Score (Hierarchical Clustering): 0.3336649836162428
Silhouette Score (K-Medoids): 0.08311783833104286
```

The combination of visual representations (scatter plots, cluster circles) and quantitative metrics (silhouette scores) offers a comprehensive evaluation of the clustering results. Together, they provide a robust basis for interpreting the geographical and demographic patterns within the fatal police shootings dataset.

Conclusion:

In summary, the analysis of fatal police shootings offers valuable insights into both demographic disparities and spatial patterns surrounding these incidents. The application of clustering algorithms, particularly KMeans, Hierarchical Clustering, and K-Medoids, contributes to a better understanding of the geographical context in which these events occur. The identification of clusters can potentially assist in uncovering patterns and trends that may inform policy decisions and interventions.

However, it is imperative to acknowledge the inherent limitations within the dataset. The analysis relies on available data, and the absence of certain key variables may limit the depth of our understanding. To draw more comprehensive conclusions and ensure a nuanced interpretation, additional contextual information about each incident, such as the circumstances leading to the encounter, socio-economic factors, and community dynamics, would be invaluable.

6. Appendix C: Data and Code

Below is the link for code:

https://github.com/bhanuprasadthota/MTH-522-Project-2/blob/main/stat2.ipynb

7. References

1. Scikit-Learn Documentation: The Scikit-Learn library was instrumental in implementing the KMeans clustering algorithm. The official documentation for Scikit-Learn provided insights into the usage and parameters of the algorithm. [Scikit-Learn Documentation](https://scikit-learn.org/stable/documentation.html)

2. Spatial Clustering Algorithms - Research Paper: To understand the theoretical foundations and concepts behind spatial clustering algorithms, we referred to the seminal work by Han, J., Kamber, M., & Pei, J. on clustering techniques for spatial data. [Han, J., Kamber, M., & Pei, J. (Year). "Data Mining: Concepts and Techniques." Publisher: Elsevier.]

3. Statistical Methods in Demographic Analysis - Book: For the demographic analysis section, we consulted the book "Statistical Methods for Demographic Research" by Agresti, A. This resource guided our approach to analyzing demographic disparities in fatal police shootings. [Agresti, A. (Year). "Statistical Methods for Demographic Research." Publisher: Wiley.]

4. Fatal Police Shootings Database: The dataset used in this project was sourced from The Washington Post's "Fatal Force" database. The database contains information on fatal police shootings, and users can access it at [The Washington Post - Fatal Force Database](https://www.washingtonpost.com/graphics/investigations/police-shootingsdatabase/)

5. Elbow Method in KMeans Clustering - Research Paper: Our decision on the optimal number of clusters using the elbow method was influenced by the research paper "A Comparative Study of K-Means and K-Medoids Algorithms" by Jain, A. K. [Jain, A. K. (Year). "A Comparative Study of K-Means and K-Medoids Algorithms." Journal: Pattern Analysis and Machine Intelligence.]

8. Contributions

Bhanu Prasad Thota took the lead in data preprocessing, handling missing values, and dropping redundant features. Additionally, I performed a comprehensive demographic analysis, including visualizations of race, armed status, manner of death, age, gender, signs of mental illness, threat level, flee status, and body camera usage. I also implemented the KMeans clustering algorithm and contributed to the interpretation of both demographic disparities and spatial patterns. My role was crucial in shaping the initial insights drawn from the dataset.

Naga Venkata Lokeswarao Maturi primary contribution was in conducting spatial analysis using Hierarchical Clustering. I explored the hierarchical relationships within the dataset and generated visual representations of spatial clusters. I played a key role in interpreting geographical patterns, identifying cluster distributions, and assessing the spatial extent of each cluster using cluster circles. Additionally, I collaborated in the discussion section, providing insights into the implications of spatial patterns on policy decisions and interventions.

Mantena Harsha Vardhan Varma focused on applying the K-Medoids clustering algorithm and evaluating the results. This involved assessing the sensitivity of the clustering to outliers, which is a notable feature of K-Medoids. I also contributed to the discussion of demographic disparities, bringing attention to how different clustering algorithms may influence the identification of these disparities. My role was crucial in providing a diverse perspective on clustering techniques and their impact on the analysis.

Lakkamraju Hitesh Kashyap Varma main contributions were in the data preprocessing stage, where I worked on handling missing values and performed label encoding for categorical variables. I also contributed to the overall analysis and findings, providing support in interpreting visualizations and insights derived from the clustering algorithms. My focus on label encoding was essential for ensuring that categorical variables could be effectively utilized in the clustering process.

Each team member faced the challenge of working with a dataset that had missing values, requiring careful decisions about how to handle them. The collaborative effort in addressing these challenges and applying various clustering algorithms allowed us to provide a comprehensive analysis of fatal police shootings. The diverse skills and contributions from each team member were pivotal in creating a well-rounded and insightful project.