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:

a. Missing Data:

The dataset initially contained missing values in crucial columns such as 'race', 'armed', 'age', 'name', 'flee', 'gender', 'longitude', and 'latitude'. Handling missing data is essential for accurate analysis, and in this case, the `dropna` method was utilized to remove rows with missing values in these specific columns. While this approach helps in maintaining data integrity, it is crucial to acknowledge potential biases introduced due to the removal of incomplete records. The impact of missing data on the overall analysis and conclusions should be considered.

b. Redundant Features:

Certain features, including 'id', 'name', 'date', 'city', and 'state', were identified as redundant for the intended clustering and analysis purposes. These features do not contribute significantly to the identification of patterns or clusters based on the selected criteria. Thus, to enhance the efficiency and relevance of subsequent analyses, these redundant features were dropped. However, it's important to document and communicate the decision to drop specific features to ensure transparency and enable reproducibility.

c. Scaling Issue:

In preparation for applying clustering algorithms, the dataset underwent scaling using `StandardScaler`. Standardization is crucial for algorithms that rely on distance measures, as it ensures that all features contribute equally. However, it's noteworthy that the scaling process was applied to the entire dataset, which includes both numerical and categorical variables. StandardScaler assumes a normal distribution and may not be appropriate for categorical variables, potentially affecting the performance of clustering algorithms. To address this, alternative scaling methods suitable for categorical variables, such as one-hot encoding or appropriate transformations, could be explored. The choice of scaling method should align with the nature of the features to avoid introducing unintended biases or inaccuracies.

Addressing these issues is essential for ensuring the robustness and reliability of the subsequent analysis, allowing for more accurate interpretations and conclusions based on the available data. Regular checks for missing data, careful feature selection, and appropriate scaling techniques contribute to the overall quality and validity of the analytical process.

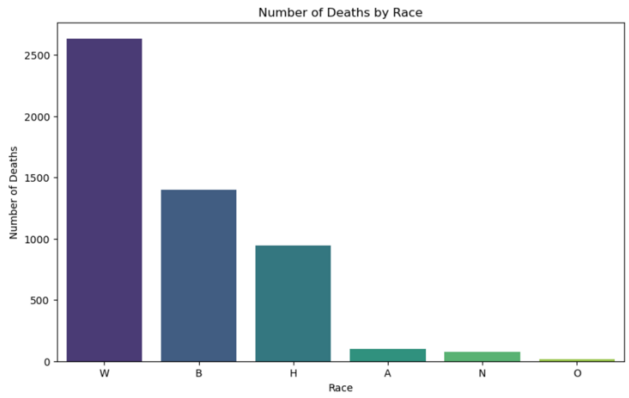
2. Findings:

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:

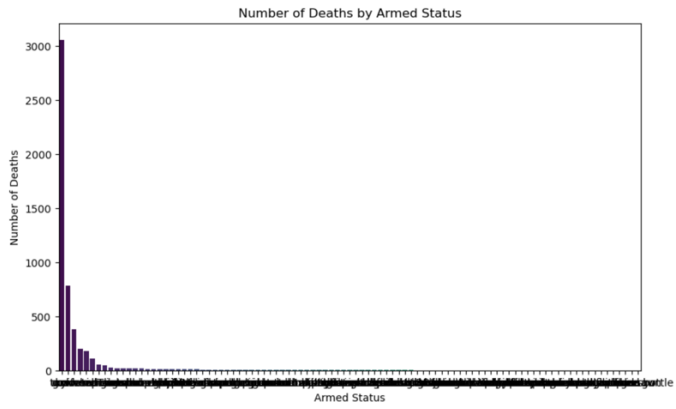
Race: The distribution of fatal police shootings across different racial groups is depicted, allowing for an examination of potential disparities.

```
Percentage of Deaths by Race:
race
W    50.878548
B    27.012937
H    18.266075
A     1.988801
N     1.486774
O     0.366866
Name: proportion, dtype: float64
========================================
```



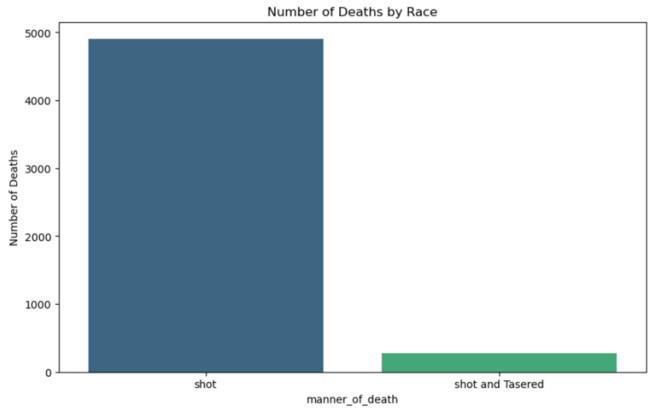
Armed Status: Insights into the proportion of incidents based on whether the individuals were armed or unarmed provide valuable context for understanding the nature of these encounters.

```
Percentage of Deaths by manner_of_death:
manner_of_death
shot                94.786638
shot and Tasered     5.213362
Name: proportion, dtype: float64
========================================
```



Manner of Death: The way fatalities occurred, whether it was through shootings or other means, is visualized to discern patterns in law enforcement actions.

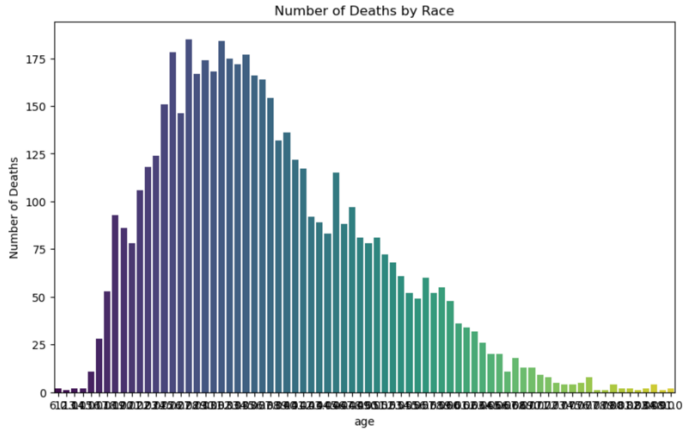
```
Percentage of Deaths by armed:
armed
gun                          58.968913
knife                        15.099440
unarmed                       7.337324
toy weapon                    3.823132
vehicle                       3.417648
                               ...
baseball bat and bottle       0.019309
fireworks                     0.019309
pen                           0.019309
chainsaw                      0.019309
flare gun                     0.019309
Name: proportion, Length: 95, dtype: float64
```



Age: Understanding the age distribution of the individuals involved in fatal encounters with the police contributes to profiling the demographics of such incidents.

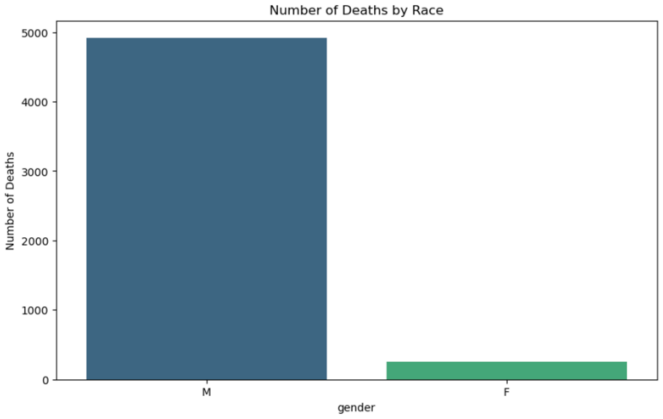
```
Percentage of Deaths by age:
age
```

```
27.0    3.572118
31.0    3.552809
25.0    3.436957
34.0    3.417648
32.0    3.379031
          ...
82.0    0.019309
12.0    0.019309
78.0    0.019309
77.0    0.019309
88.0    0.019309
Name: proportion, Length: 76, dtype: float64
```



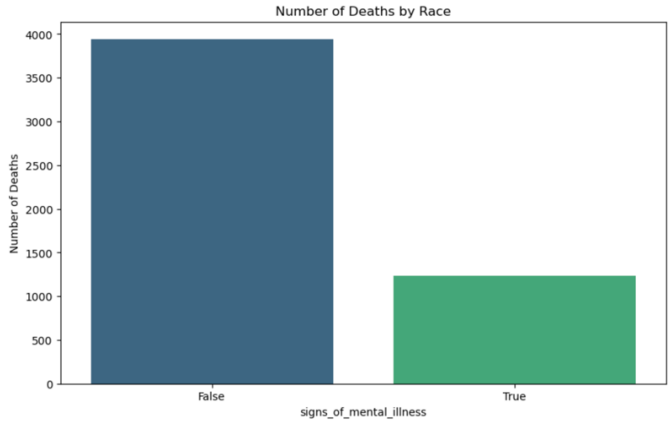
Gender: Examining the gender distribution provides insights into whether there are gender-based disparities in fatal police shootings.

```
Percentage of Deaths by gender:
gender
M    95.018343
F     4.981657
Name: proportion, dtype: float64
========================================
```

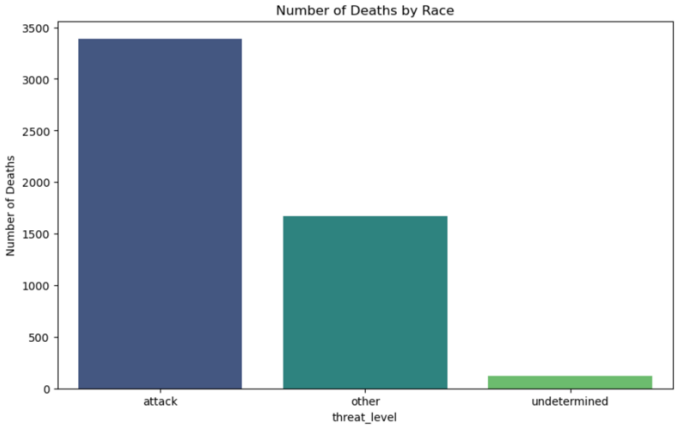


Signs of Mental Illness: The presence of signs of mental illness is explored to understand the potential role of mental health in these incidents.

```
Percentage of Deaths by signs_of_mental_illness:
signs_of_mental_illness
False    76.173006
True     23.826994
Name: proportion, dtype: float64
========================================
```



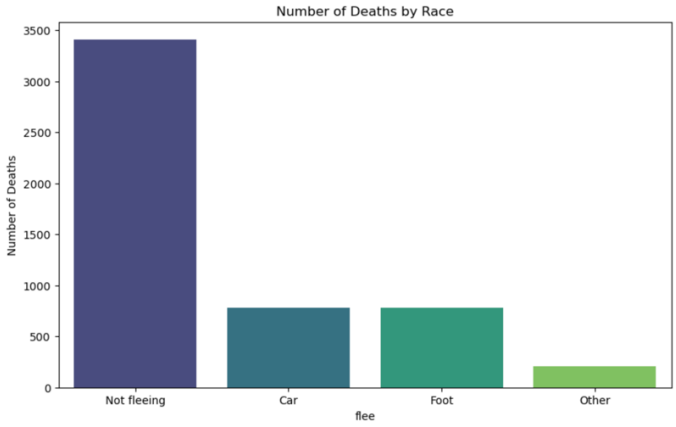
Threat Level: Analyzing the threat level associated with each incident helps in gauging the perceived threat that law enforcement faced.




```
Percentage of Deaths by threat_level:
threat_level
attack          65.495269
other           32.168372
undetermined     2.336358
Name: proportion, dtype: float64
========================================
```

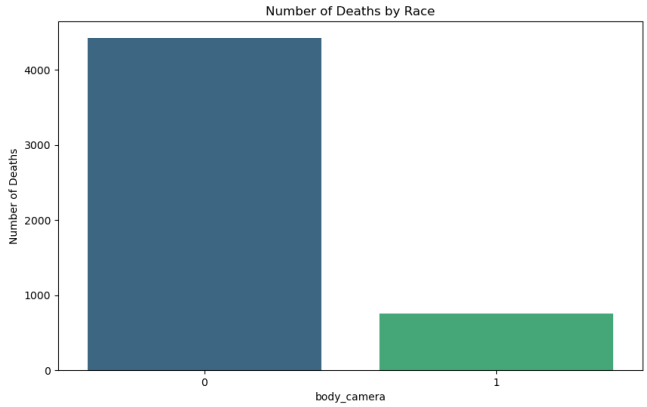
Flee Status: Understanding whether individuals were attempting to flee at the time of the incident contributes to the context of these encounters.

```
Percentage of Deaths by flee:
flee
Not fleeing    65.881444
Car            15.060823
Foot           15.022205
Other           4.035528
Name: proportion, dtype: float64
========================================
```



Body Camera Presence: The use of body cameras by law enforcement during these incidents is visualized to assess the availability of video evidence.

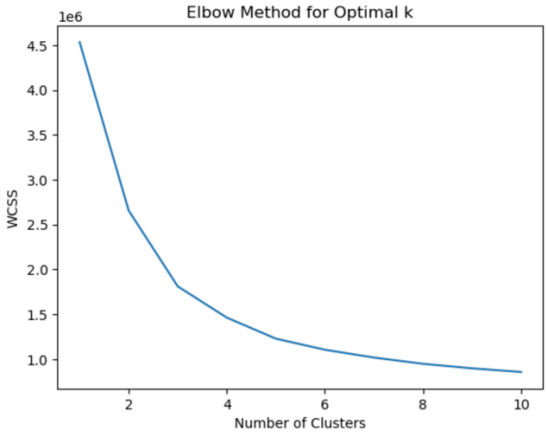
```
Percentage of Deaths by body_camera:
body_camera
False    85.402587
True     14.597413
Name: proportion, dtype: float64
========================================
```



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: The elbow method is employed to determine the optimal number of clusters in KMeans, providing a basis for the subsequent clustering analysis.

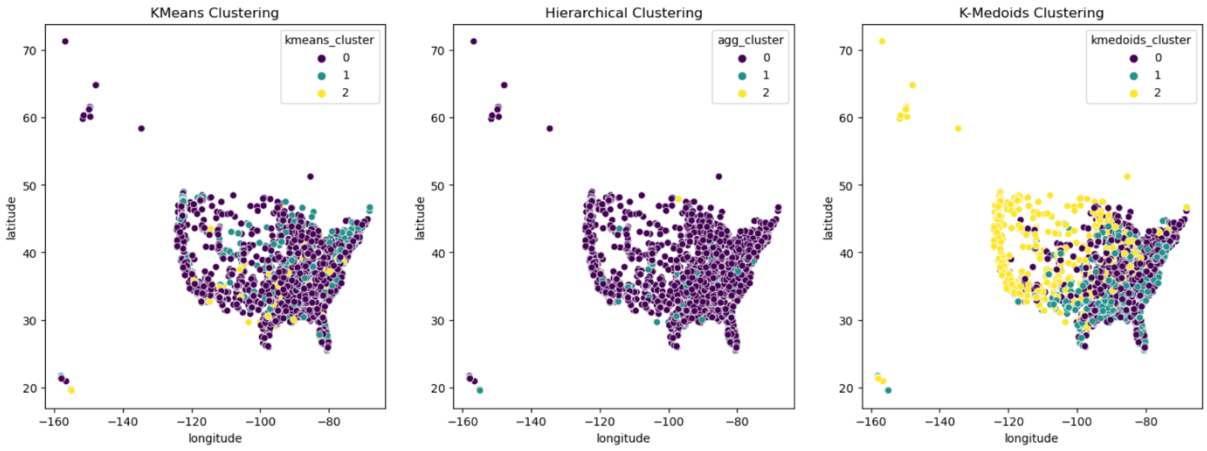


Silhouette Scores: Silhouette scores are computed to evaluate the quality of clusters produced by each algorithm. Higher silhouette scores indicate better-defined clusters.

```
Silhouette Score (KMeans): 0.16134790469330992
Silhouette Score (Hierarchical Clustering): 0.3336649836162428
Silhouette Score (K-Medoids): 0.08311783833104286
```

c. Cluster Visualization:

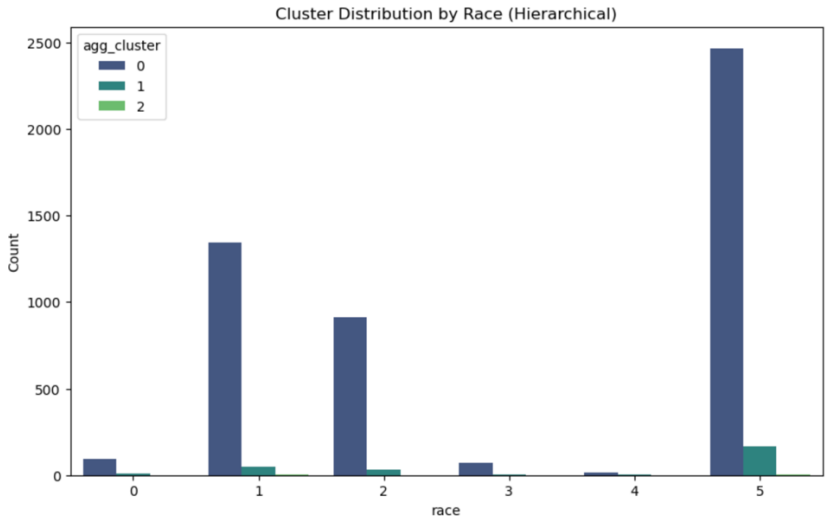
To provide a geographical perspective, scatter plots are generated to visually represent the spatial distribution of clusters. Additionally, circles are plotted around cluster centers to highlight the spatial extent of each cluster.

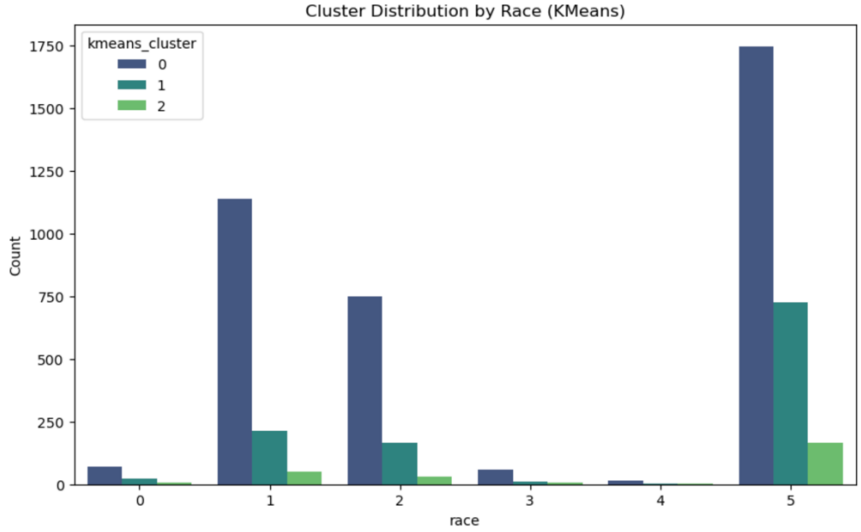


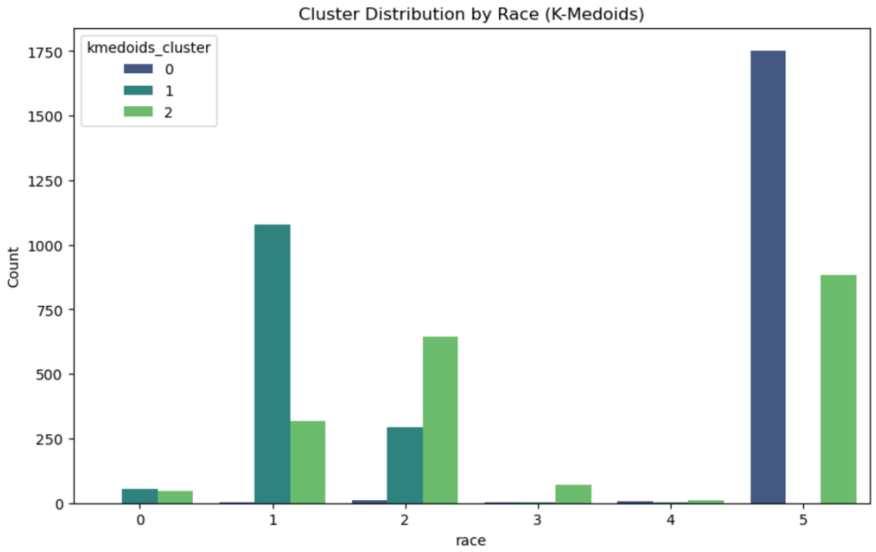
d. Cluster Distribution:

The distribution of race within each cluster is examined to identify potential patterns or disparities in geographical clusters. This analysis contributes to understanding whether certain racial groups are disproportionately affected in specific geographic areas.

These findings collectively contribute to a multifaceted understanding of fatal police shootings, encompassing demographic insights, spatial patterns, and potential disparities within clusters. The integration of demographic analysis with spatial clustering enhances the depth of interpretation and facilitates a more nuanced exploration of the dataset.







3. Discussion:

a. Demographic Disparities:

The demographic analysis uncovered noteworthy disparities in the distribution of fatalities, particularly concerning race, gender, and age. The bar plots depicting the number of deaths across different demographic categories provide a snapshot of the unequal impact of fatal police shootings on various groups. For instance, understanding the disproportionate representation of specific racial or age groups among the victims is crucial for addressing potential systemic issues. However, these findings represent associations and not causations, emphasizing the need for a deeper investigation to comprehend the root causes of these disparities.

Moreover, delving into the intersectionality of these demographic factors can provide a more nuanced understanding. Intersectionality considers how different aspects of identity, such as race, gender, and age, intersect and interact, potentially contributing to unique experiences and vulnerabilities. Further research and analysis should explore these intersections to unravel complex relationships within the data.

b. Spatial Patterns:

The utilization of clustering algorithms revealed distinct spatial patterns in fatal police shootings. The scatter plots, overlaid with clustered data points and circles, offer a visual representation of geographical concentrations. Identifying these spatial patterns is a crucial step toward understanding the context in which these incidents occur. It prompts questions about the relationship between geographical factors and the occurrence of fatal police shootings. For instance, are certain areas more prone to such incidents due to socioeconomic factors, community dynamics, or policing practices? Interpretation of these clusters should consider not only the geographic locations but also the social and economic context of the regions in question.

Examining spatial patterns can guide policymakers and law enforcement agencies in implementing targeted interventions to address specific challenges in high-incidence areas. Additionally, it facilitates a more informed dialogue on the allocation of resources and the development of community-based strategies to reduce instances of fatal police shootings.

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:

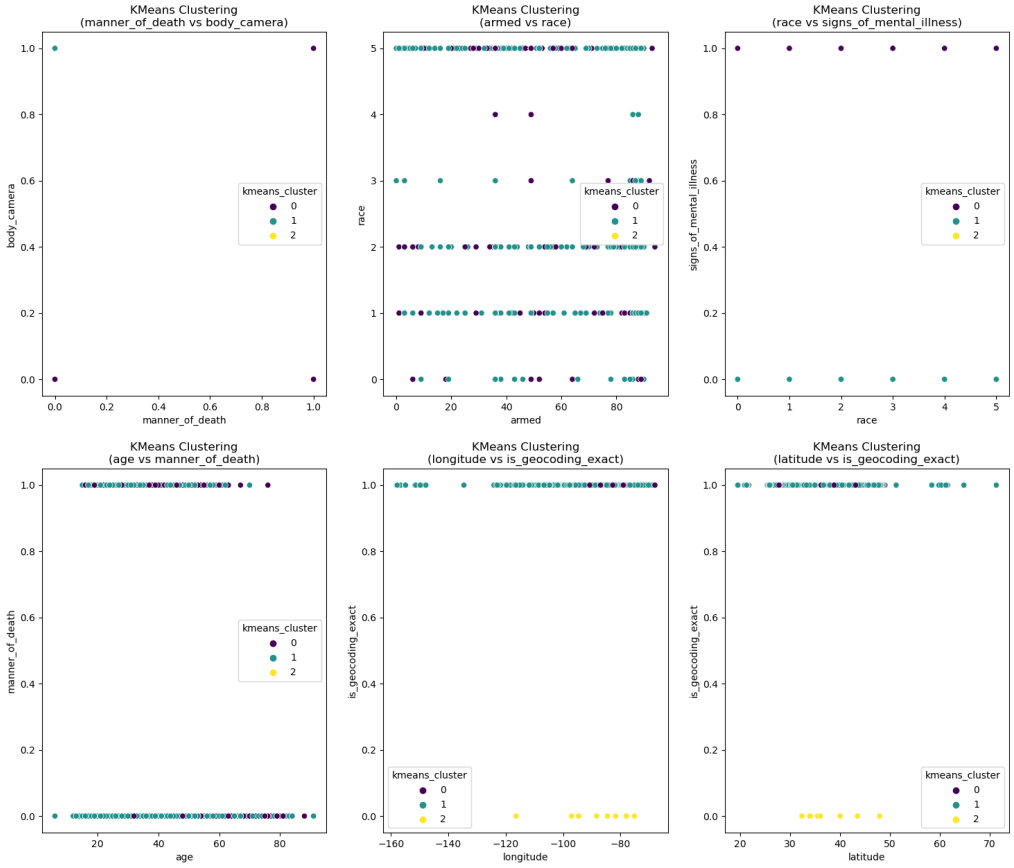
The dataset used in this analysis is obtained from an Excel file that compiles information on fatal police shootings. The Excel file, located at the specified path "D:\MTH\PROJECT2\fatal-police-shootings-data.xls," serves as the primary source for the investigation. The dataset likely includes various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

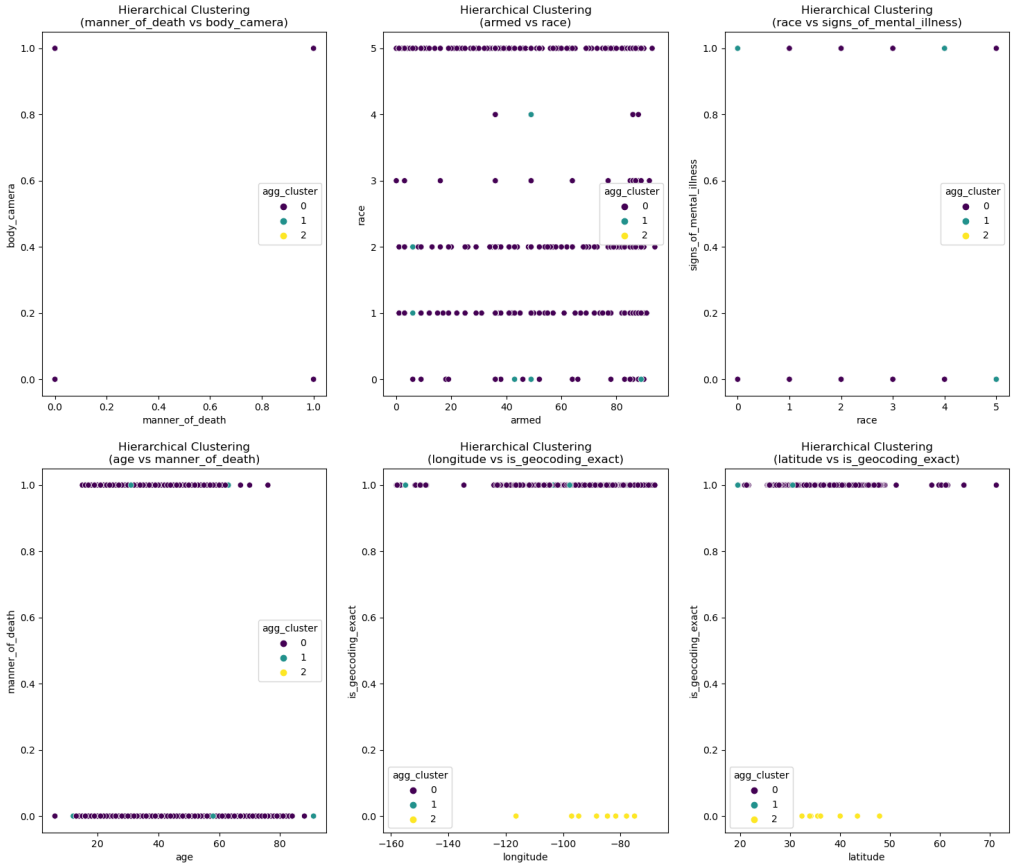
3. Analytic Methods:

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).



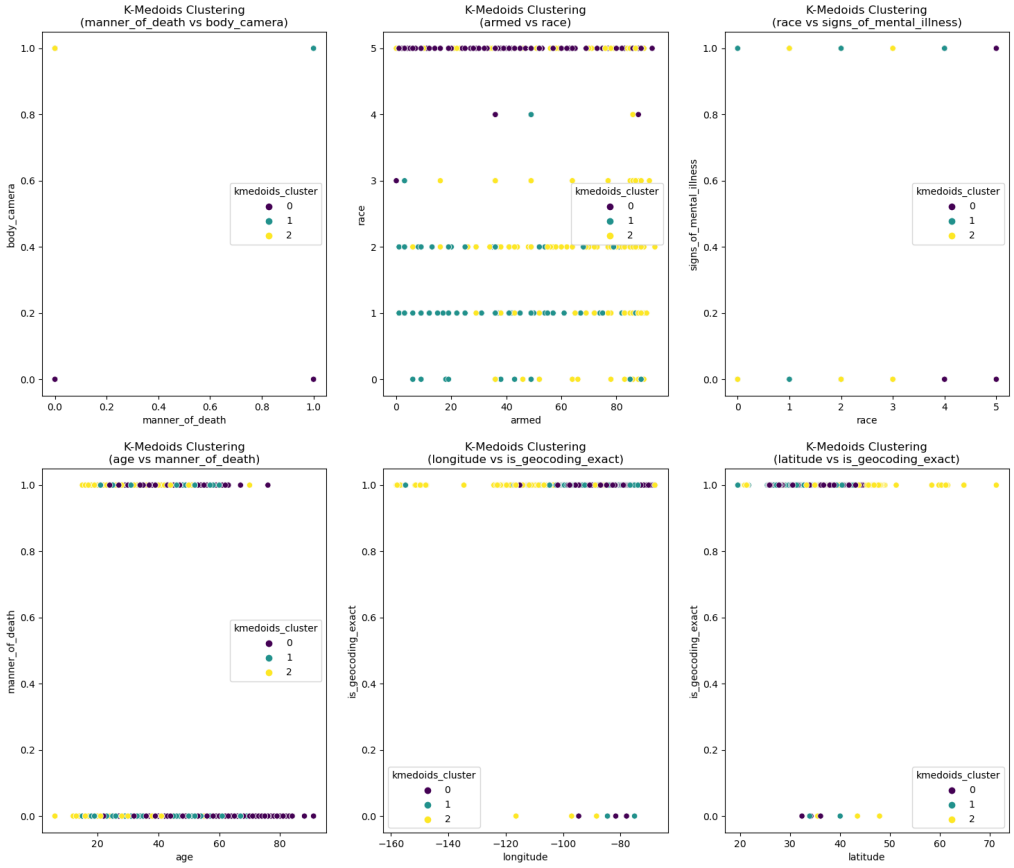
b. Hierarchical Clustering:

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, K-Medoids 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.

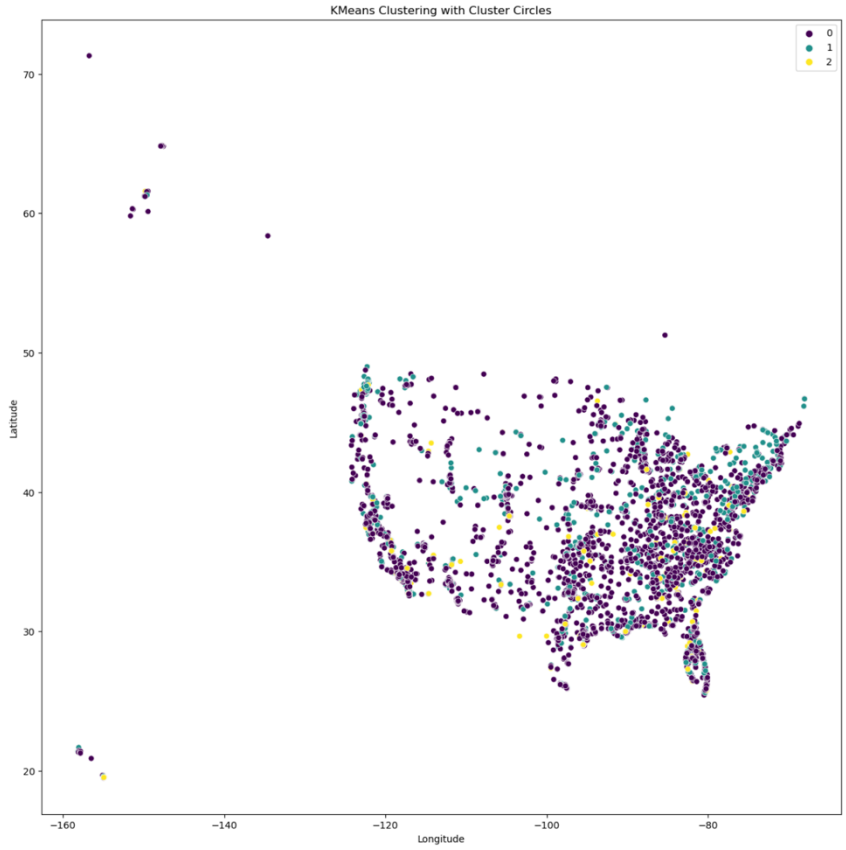
5. Appendix B: Results

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-shootings-database/)

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.