Unveiling Spatial Trends: K-Means Analysis of Police Shootings in US

Authors: Veda Sahaja Bandi, Sindhuja Baikadi

The Issues:

In this analysis, we focus on applying clustering techniques to a dataset that documents fatal police shootings across the continental United States, aiming to gain deeper insights into the matter. The dataset itself is extensive, yet it poses challenges due to missing values, particularly in critical variables like longitude and latitude, which are fundamental for spatial analysis. We aim to uncover potential patterns or clusters within the geographical distribution of these incidents across the nation.

As we embark on the exploratory aspect of our analysis, a multitude of questions naturally come to mind, prompting us to delve deeper into the dataset and its implications.

- How do we handle missing geographical coordinates, especially in variables crucial for spatial analysis?
- Determining the optimal number of clusters to use in this evaluation, and how to validate this choice effectively?
- Can the clustering results offer meaningful insights into the distribution of police shootings? If so, what implications might these patterns hold regarding underlying societal or systemic factors?

Addressing these inquiries is crucial for unveiling potential spatial trends in the data and laying the groundwork for deeper exploration into the dynamics of Police Shootings in the United States.

Findings:

Applying a clustering algorithm to the dataset of fatal police shootings in the continental United States revealed clear groupings that correspond to specific geographical regions. Despite encountering missing data, especially in crucial variables for spatial analysis like geographical coordinates, the analysis successfully identifies significant clusters when visualized on a map. The decision to remove the corresponding data points from the dataset appears to be a prudent approach.

The clustering results reveal a notable spatial grouping of fatal police shooting incidents across the continental United States, indicating potential regional hotspots. The analysis determines an optimal number of three clusters, suggesting distinct geographical areas for categorizing police shooting incidents.

Visualizing the data by color-coding clusters reveals frequent police shootings in both the Eastern and Western regions of the United States, with major cities like New York, Houston, and Los Angeles exhibiting concentrated clusters. This concentration suggests a significant portion of police shootings occur within urban areas, emphasizing the prevalence within densely populated cities. The statistical variance within clusters indicates that there's a lot of variation within the clusters arguing that the incidents are spread out across different areas, even though they're grouped by location.

These findings underscore the necessity for further investigation into law enforcement practices, socio-economic conditions, and community relations in urban environments, offering insights crucial for understanding the dynamics of police shootings and guiding policy reforms to mitigate such incidents and address systemic issues.

Discussion:

The analysis reveals spatial patterns in fatal police shootings, suggesting potential regional disparities influenced by factors like population density, law enforcement practices, and socio-economic variables. It is apparent that the shootings are not randomly distributed but are instead concentrated in specific areas. This raises crucial questions about the underlying factors shaping the distribution of police shootings across the United States.

The pronounced clustering indicates potential correlations with external variables such as urbanization, crime rates, socio-economic status, and population demographics. Understanding whether these clusters are associated with specific socio-economic or cultural contexts requires further investigation. Additionally, the high variance within clusters suggests disparities in the spread of shootings within each region, hinting at the presence of distinct subregions or neighborhoods with unique characteristics.

Further research is needed to unravel the complex interplay of variables contributing to these patterns, including law enforcement protocols, community policing effectiveness, regional gun laws, and historical socio-political contexts. However, it's essential to acknowledge that the unsupervised nature of the algorithm limits its ability to assign causation or offer deeper insights into the underlying reasons for these spatial patterns.

Appendix A: Method

Data was downloaded as a comma-separated (.csv) file and imported into Jupyter Notebook. The analysis was done on the dataset which consisted of the longitude and latitude of the geolocations in the United States where fatal Police Shootings have taken place.

Data cleaning procedures were then executed to address missing values, particularly in the 'latitude' and 'longitude' columns which are crucial for spatial analysis. Records lacking geographic coordinates were excluded to ensure the accuracy of subsequent clustering processes.

The K-Means clustering algorithm from the sci-kit-learn library was used to identify patterns in the geographic distribution of police shootings. The optimal number of clusters was determined using the Elbow Method, which involved plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters and iterating the algorithm over a range of cluster numbers (1 to 10) to record the inertia (WCSS) for each iteration.

The K-Means Clustering was performed on the data with the number of clusters set to three based on the Elbow Method. To visualize the results, Folium, a Python library tailored for geospatial data visualization, was employed to create an interactive map. Data points representing clustered incidents were differentiated by color on the map, facilitating a clear visual representation of their geographic distribution and density.

This systematic approach enabled a structured analysis of the dataset, effectively exploring and visualizing the spatial patterns of fatal police shootings. Relevant Python code can be found in Appendix C.

Appendix B: Results

The CSV File contains 7912 data points containing the data for fatal Police shootings in the United States along with the latitude and longitude of the locations. Out of this, 827 rows of null values are removed from the dataset.

The analysis employed the K-Means clustering algorithm from the sci-kit-learn library to discern patterns in the geographic distribution of police shootings. The optimal number of clusters was determined through the application of the Elbow Method.

The Elbow Method plot (Figure 1) indicates that three clusters provide a reasonable balance between the within-cluster sum of squares (WCSS) and the number of clusters, suggesting that the chosen number of clusters is appropriate for the given data.

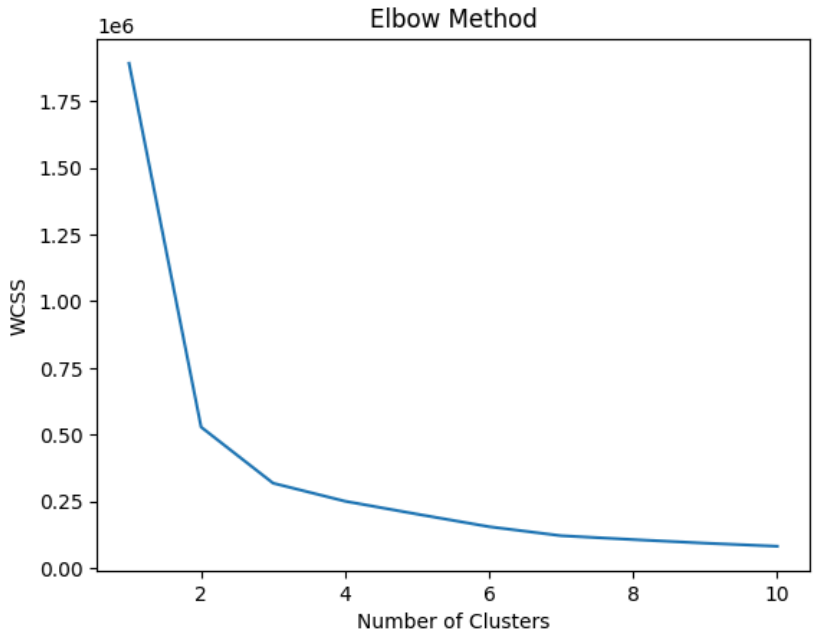

Figure 1: Elbow Method

The K-Means algorithm identified three distinct clusters within the dataset, which represent different geographical regions of the United States.

These clusters were visualized on a map, with each cluster represented by different colors: blue (Cluster 1), red (Cluster 2), and green (Cluster 3). These clusters appear to align with specific regions, suggesting potential geographical correlations in the occurrence of these incidents.

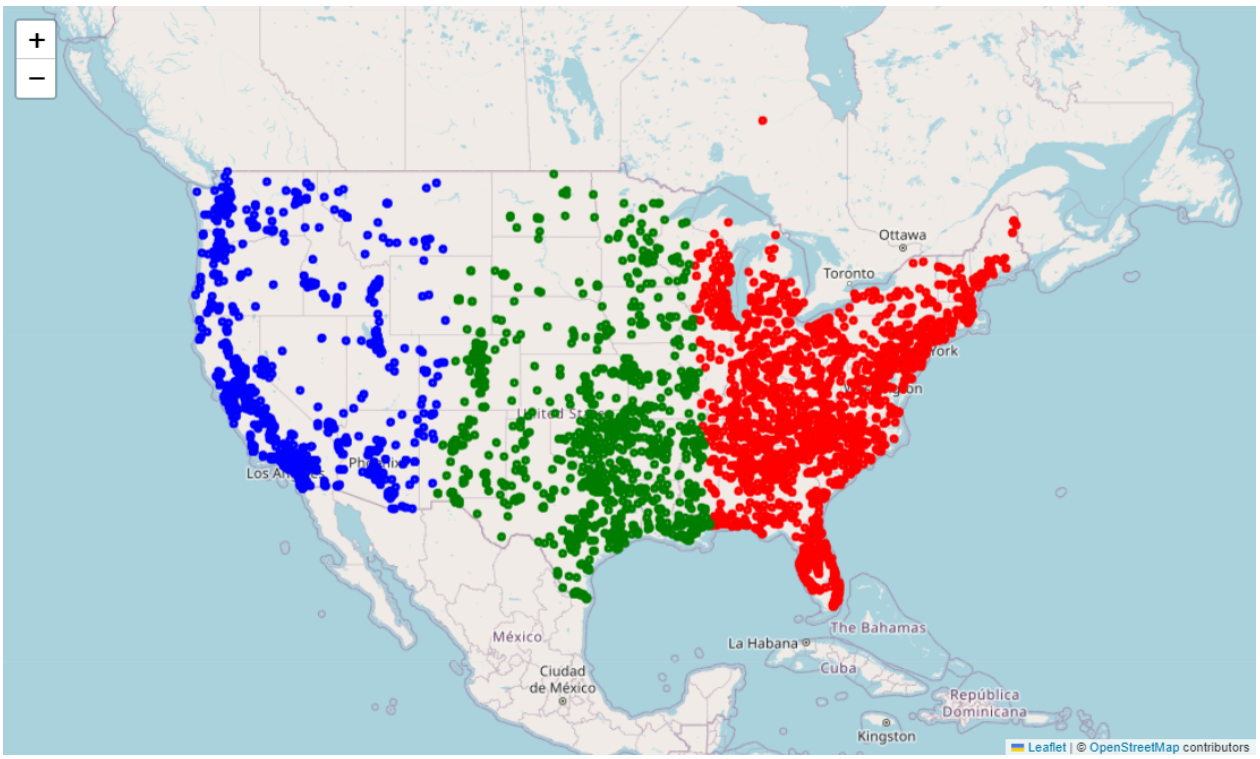

Figure 2: Interactive map with the clusters

The map visualization, enriched by the color-coded clusters, provides a comprehensive depiction of the geographical distribution of incidents across the United States. Notably, there are pronounced concentrations of incidents observed along the East Coast, West Coast, and Southern regions. Such spatial clustering hints at potential regional disparities in the frequency of police shootings, which could be attributed to diverse factors including population dynamics, levels of urbanization, and variations in policing strategies and practices throughout the nation.

The inertia value of 312880 from the K-Means clustering indicates a degree of cohesion within the clusters, although the spread within each cluster suggests that the incidents are geographically dispersed rather than tightly localized.

The initial data examination revealed missing values in key geographical coordinates. However, the filtered dataset used for clustering retained a substantial number of incidents, ensuring the analysis remained robust and relevant.

The visualization of the data points on the interactive map (Figure 3) provides insight into the spatial distribution of police shootings, showcasing clusters predominantly situated in regions characterized by higher population densities. This observation underscores the correlation between population density and the occurrence of police shootings, suggesting that areas with denser populations are more likely to experience such incidents.

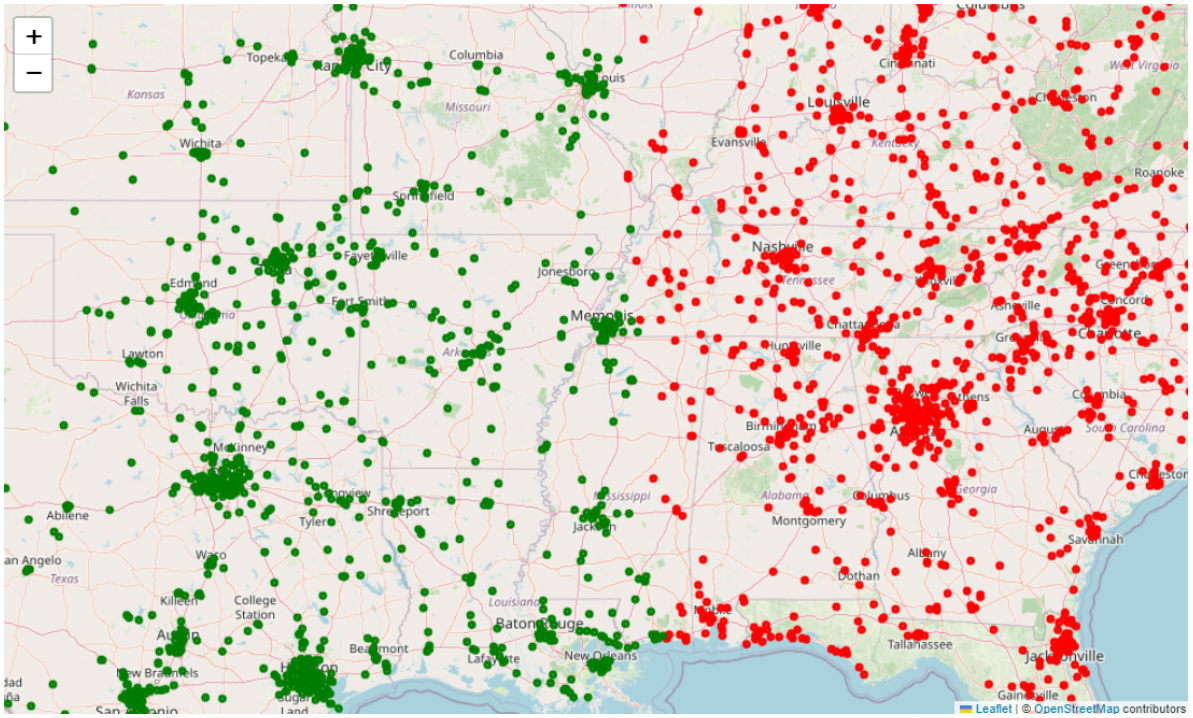

Figure 3: Interactive map with the clusters located in Urban Locations

The results of the clustering process, while revealing distinct spatial groupings, do not provide direct insights into the causative factors of police shootings. The observed patterns necessitate further investigation into socio-economic, demographic, and law enforcement policies to understand the underlying causes of these regional trends.

In summary, the clustering analysis offers a foundational understanding of the spatial characteristics of fatal police shootings across the continental United States. These results serve as a stepping stone for more detailed investigations into the factors influencing these patterns, with implications for policy formulation, law enforcement training, and community relations.

Appendix C: Code

In this appendix, we document the Python code for performing K-Means Clustering on the Police Shooting in the continental United States.

```
# Importing the libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import folium
from geopy.distance import geodesic
```

```
from sklearn.cluster import KMeans
```

```
# K-Means Clustering
model = KMeans(n_clusters=3)
y_kmeans = model.fit_predict(df)
df = pd.concat([df, pd.DataFrame(y_kmeans, columns=["y"])], axis=1)
model.inertia_
```

Output

312880.1429418013

```
# Elbow Method and Plot
wcss = []
for i in range(1,11):
    model = KMeans(n_clusters=i, n_init=10)
    y_kmeans = model.fit_predict(df)
    wcss.append(model.inertia_)
plt.plot(range(1,11), wcss)
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.title('Elbow Method')
plt.show()
```

```
# Interactive Map with K-Means Clustering
cluster1 = df[['latitude','longitude']][df['y']==0].values.tolist()
cluster2 = df[['latitude','longitude']][df['y']==1].values.tolist()
cluster3 = df[['latitude','longitude']][df['y']==2].values.tolist()
map = folium.Map(location=[data['latitude'].iloc[0],data['longitude'].iloc[0]], zoom_start = 10, tiles="openstreetmap")
for i in cluster1:
    folium.CircleMarker(i, radius=2, color='blue', fill_color='lightblue').add_to(map)
for i in cluster2:
    folium.CircleMarker(i, radius=2, color='red', fill_color='lightred').add_to(map)
for i in cluster3:
    folium.CircleMarker(i, radius=2, color='green', fill_color='lightgreen').add_to(map)
```

map

Contribution:

Sindhuja Baikadi - 02128756: Worked on the Issues, Findings, Discussion, Method, and Results sections. Also self-plotted the graphs to analyze the data using the various methods discussed in the report.

Veda Sahaja Bandi - 02105111: Outlining key observations and insights, worked on the coding portion of the project to implement necessary functionalities and features.