**A close-up of a logo

Description automatically generated**

**Advanced Mathematical**

**Statistics (MTH 522)**

**Project 3**

**Unveiling Patterns of Political Violence: A K-Means Clustering Analysis of India's Geopolitical Landscape**

**Submitted By:**

**Supreeth Mohan - 02036259**

**Roshni Pal - 02137180**

**Trina Xavier - 02102403**

**Aryan Bhalla – 02107402**

**The Issues:**

In this project, Our primary objective was to uncover comprehensive analysis of a dataset aimed at understanding the patterns of political violence incidents in India. By employing clustering techniques, we sought to identify geographic hotspots, temporal trends, and demographic factors associated with political violence, for a deeper understanding of the complex dynamics for unrest in the region.

Below are the primary issues we are examining:

1. Determine if there are top 5 specific regions or areas in India that are more prone to political violence than others.
2. Are there spatial trends or regional concentrations of political violence incidents? Does the cluster analysis reveal specific hotspots or regions with heightened political tension?
3. Investigating if there is a correlation between different variables such as the number of fatalities, event types, and geographical locations of violence incidents.

**Findings:**

The k-means clustering algorithm is an unsupervised machine learning technique that groups data points into a specific number of clusters (k). In our case, the data points are the **individual incidents of political violence**, and the clusters are groups of incidents that are geographically close to each other. Here are some key findings from our analysis:-

* The analysis reveals insights for six clusters based on their incident counts in descending order:

1. Cluster 0 has 10,397 incidents.

2. Cluster 2 has 8,164 incidents.

3. Cluster 3 has 7,059 incidents.

4. Cluster 5 has 4,800 incidents.

5. Cluster 4 has 3,042 incidents.

6. Cluster 1 has 2,921 incidents.

* The findings indicate that **Cluster 0** is the most violent, with **10,397 incidents**, while **Cluster 1** is the least violent, with **2,921 incidents**. This analysis underscores the varying levels of political violence across the different clusters, with Cluster 2 identified as the second most violent and Cluster 4 as the least violent.
* The top five repeated latitudes identified in the cluster analysis provide key geographic insights into the concentration of political violence incidents in India. Specifically, **latitude 22.5728, with 500 incidents**, stands out as the most frequently occurring latitude, indicating a significant hotspot of political violence.
* The analysis reveals that the longitude **88.3747** exhibits the highest frequency of political violence incidents, with a total of **513 occurrence**s. This longitude represents a significant hotspot for political unrest and conflicts.
* The frequent occurrence of 500 incidents at the coordinates 22.5728 latitude and 88.3747 longitude highlights a significant vulnerability to political violence in this specific location. This vulnerability may be influenced by various factors, including historical tensions, socio-economic disparities, or ethnic complexities.
* The top five repeated locations identified in the cluster analysis provide significant insights into the distribution of political violence incidents across India:

**Top 5 Repeated Locations:**

* 1. Kolkata, West Bengal
  2. Aizawl, Mizoram
  3. Silchar, Assam
  4. New Delhi, Delhi
  5. Warangal, Telangana
* After conducting cluster analysis to investigate potential correlations between variables such as the **number of fatalities, event types, and geographical locations** of violence incidents, the **top 5 events** within each cluster revealed distinct patterns:
* **Cluster 1** predominantly comprised **violence against civilians**.
* **Cluster 2** was **battles**.
* **Cluster 3** **violence against civilians**.
* **Cluster 4** predominantly **battles**.
* **Cluster 5** primarily consisted of **Riots**.

**Discussion:**

Based on the findings from the statistical analysis, we can have the following discussions:-

1. It is important to note that k-means clustering is a descriptive technique, not an inferential technique. This means that it can identify patterns in the data, but it cannot be used to draw causal conclusions about why the patterns exist. For example, the k-means clustering cannot tell us why there is a higher concentration of political violence in Cluster 0 compared to the other clusters.
2. The k-means clustering algorithm assumes that the data is evenly distributed in space. However, political violence incidents may not be evenly distributed, and this could affect the results of the clustering analysis.

**Appendix A: METHOD**

We sourced the Locations of Political Violence in India Dataset from the provided class link, importing it into a Jupyter Notebook. In our analysis, we amalgamated the ***'age' and 'race' columns,*** leading to an exploration of the intricate relationship between them.

1. **Data Collection:**

The data used in this study was obtained from The Locations of Political Violence in India Dataset provided below.

<https://www.dropbox.com/scl/fi/rnhn88bfzbo26ugq408rj/IndiaPoliticalViolence-no-Islands_Kashmir_Ladakh.csv?rlkey=9676jgcbhjb2kv9azkg998jiv&e=1&dl=0>

1. **Data Preparation:**

The dataset was downloaded and examined; the procedure was documented.

1. **Variable Creation:**

Using Pandas, we retrieved the .csv file– IndiaPoliticalViolence(no Islands\_Kashmir\_Ladakh).csv and loaded it to the variable called data.

In that .csv file, the data *set* ***has age and race, columns then from the race column data for black people is extracted as black\_ages, and the same goes for white as white\_ages.***

**Analytic Method:**

The statistical procedures used in the above analysis are as follows:

**Data Preprocessing:**

Cleaning and preprocessing of the dataset to ensure consistency and reliability.

Handling missing values, outliers, and irrelevant variables.

**Descriptive statistics:** Calculation of summary statistics such as counts, means, and standard deviations to understand the distribution of political violence incidents.

**K-means clustering:** Application of the K-means algorithm to partition the dataset into distinct clusters based on similarity in features such as geographical coordinates, type of violence, and severity of incidents.

**Cluster visualization:** Visualization of clustered data using scatter plots, heatmaps, or geographic maps to identify spatial patterns and clusters of political violence incidents.

**APPENDIX B: RESULT**

**Geographical Distribution of Political Violence in India: Two Distinct Clusters Identified**

The k-means clustering analysis has identified two distinct clusters of political violence incidents in India. Here's a breakdown of the information:

1. Cluster 0 (Orange): This cluster has the highest number of incidents (17,146) and is likely the most concentrated area of political violence in India.
2. Cluster 1 (Blue): This cluster has 10,729 incidents.

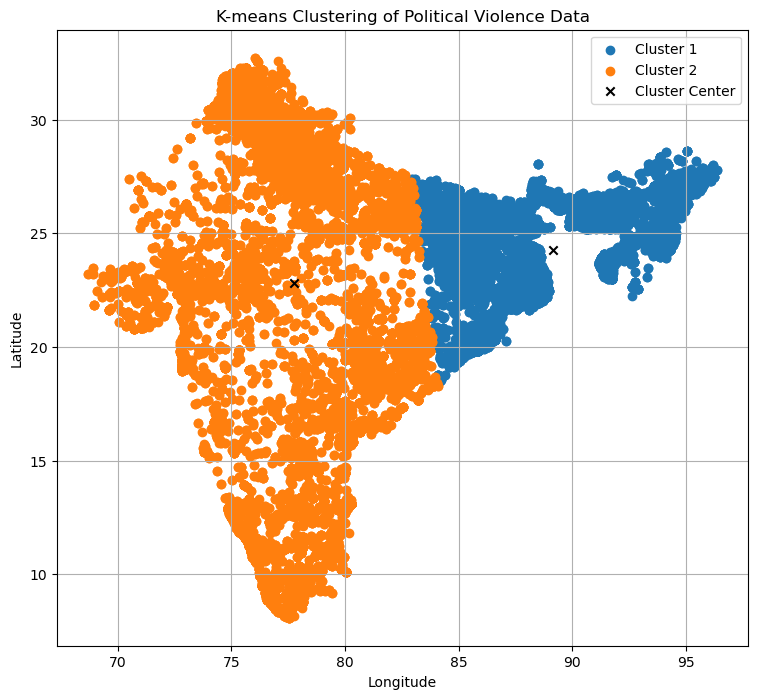
****

Fig 1. Geographical Distribution of Political Violence in India: Two Distinct Clusters Identified

**K-Means Clustering Results: 3 Clusters of Political Violence Incidents in India:**

The k-means clustering analysis has identified three distinct clusters of political violence incidents in India. Here's a breakdown of the information:

1. Cluster 0 (Red): This cluster has the highest number of incidents (17,146) and is likely the most concentrated area of political violence in India.
2. Cluster 1 (Blue): This cluster has 10,729 incidents.
3. Cluster 2 (Green): This cluster has the least number of incidents (8,508)

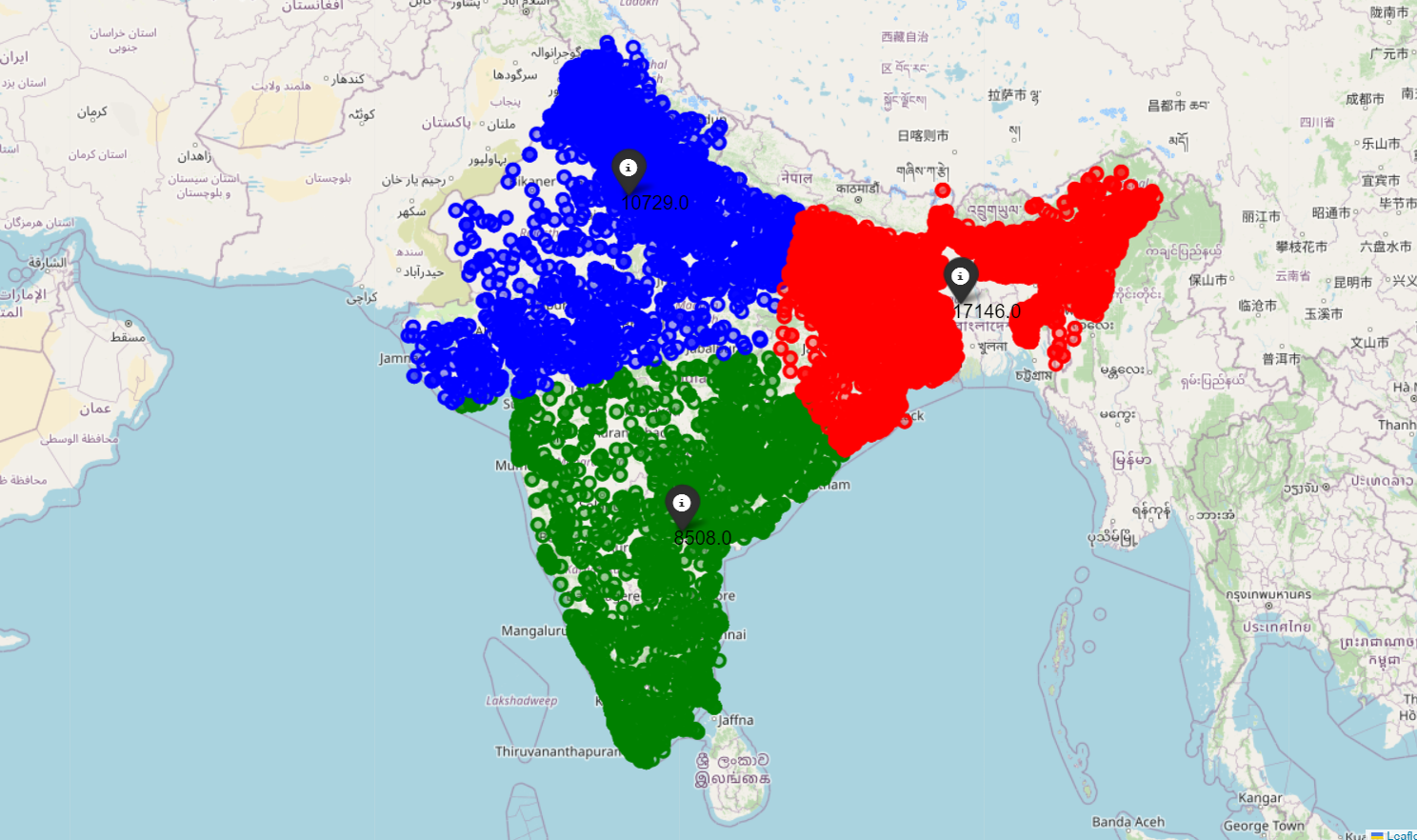


Fig 2. K-Means Clustering Results: 3 Clusters of Political Violence Incidents in India

The below 3 clusters suggest three geographically distinct areas of political violence in India, with **Cluster 0 being the most concentrated.** However, further analysis is needed to understand the underlying reasons behind these patterns.

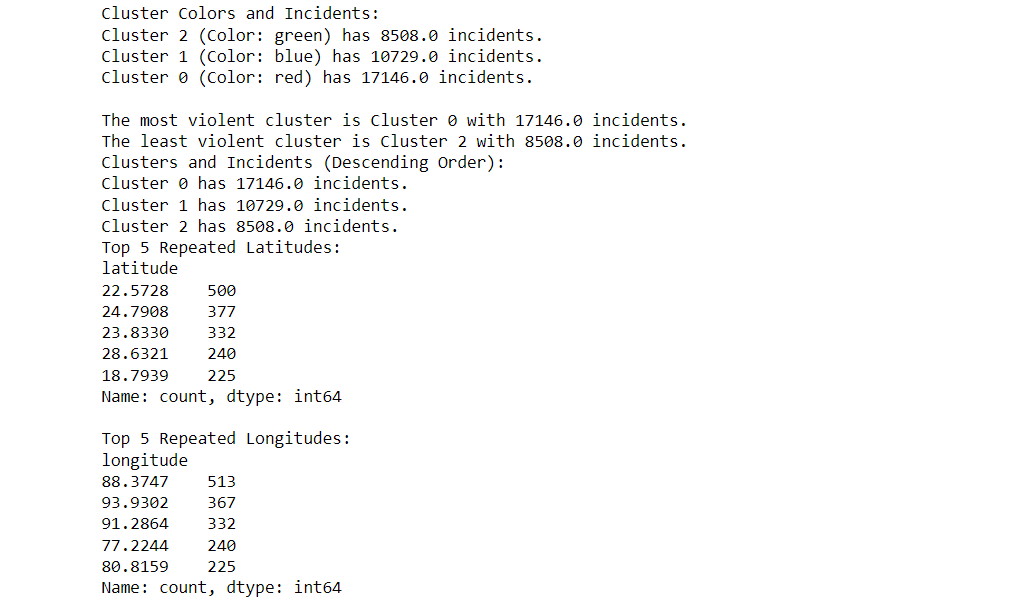


Fig 3. K-means clustering description for 3 clusters

**K-Means Clustering Results: Identifying 4 Geographically Distinct Clusters of Political Violence in India:**

The k-means clustering analysis identified four distinct clusters of political violence incidents in India. The clusters are represented by different colors (e.g., red, blue, green, yellow) on the map.

Cluster 0 (orange): This cluster has the highest number of incidents (11,162) and is likely the most concentrated area of political violence in India.

Cluster 3 (Red): This cluster has 10,080 incidents.

Cluster 2 (Blue): This cluster has 8,011 incidents.

Cluster 3 (Green): This cluster has the least number of incident 7130

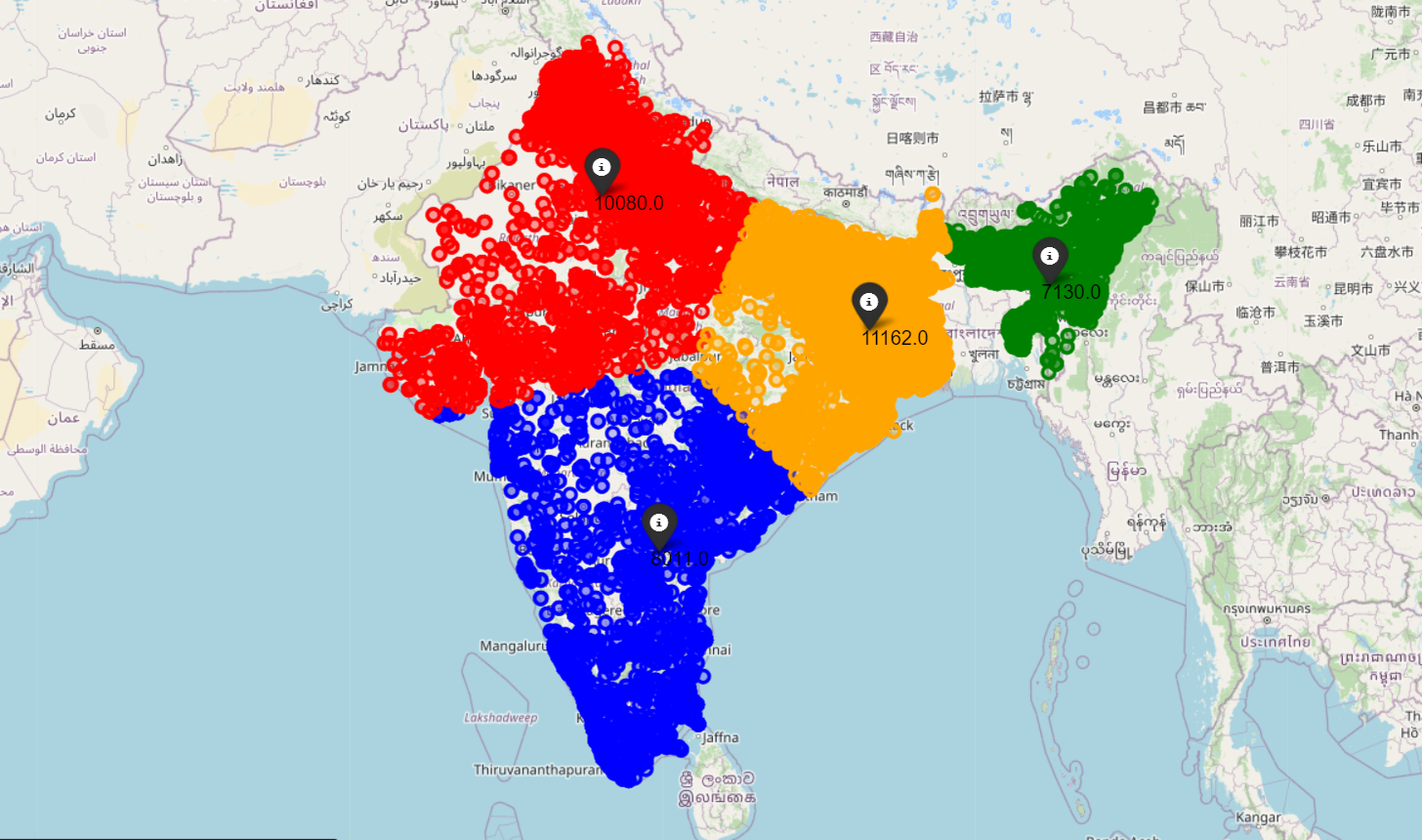


Fig 4. K-Means Clustering Results: Identifying Four Clusters

The results of the cluster analysis show that **Cluster 0 (orange) has the most incidents (11,162.0),** followed by Cluster 3 (red) with 10,080.0 incidents, Cluster 2 (blue) with 8,011.0 incidents, and Cluster 1 (green) with the least incidents at 7,130.



Fig 5. K-means clustering description for 4 clusters

**K-Means Clustering Revealing 5 Distinct clusters of Political Violence in India:**

The k-means clustering analysis identified five distinct clusters of political violence incidents in India. Here's a breakdown of the information:

Cluster 2 (Purple): This cluster has the highest number of incidents (10,856) and is likely the most concentrated area of political violence in India.

Cluster 3 (Orange): This cluster has (8,607 )incidents.

Cluster 1 (Red): This cluster has (7,114) incidents.

Cluster 0 (Blue): This cluster has (7,130) incidents.

Cluster 4 (Green): This cluster has the least number of incidents (2,996)

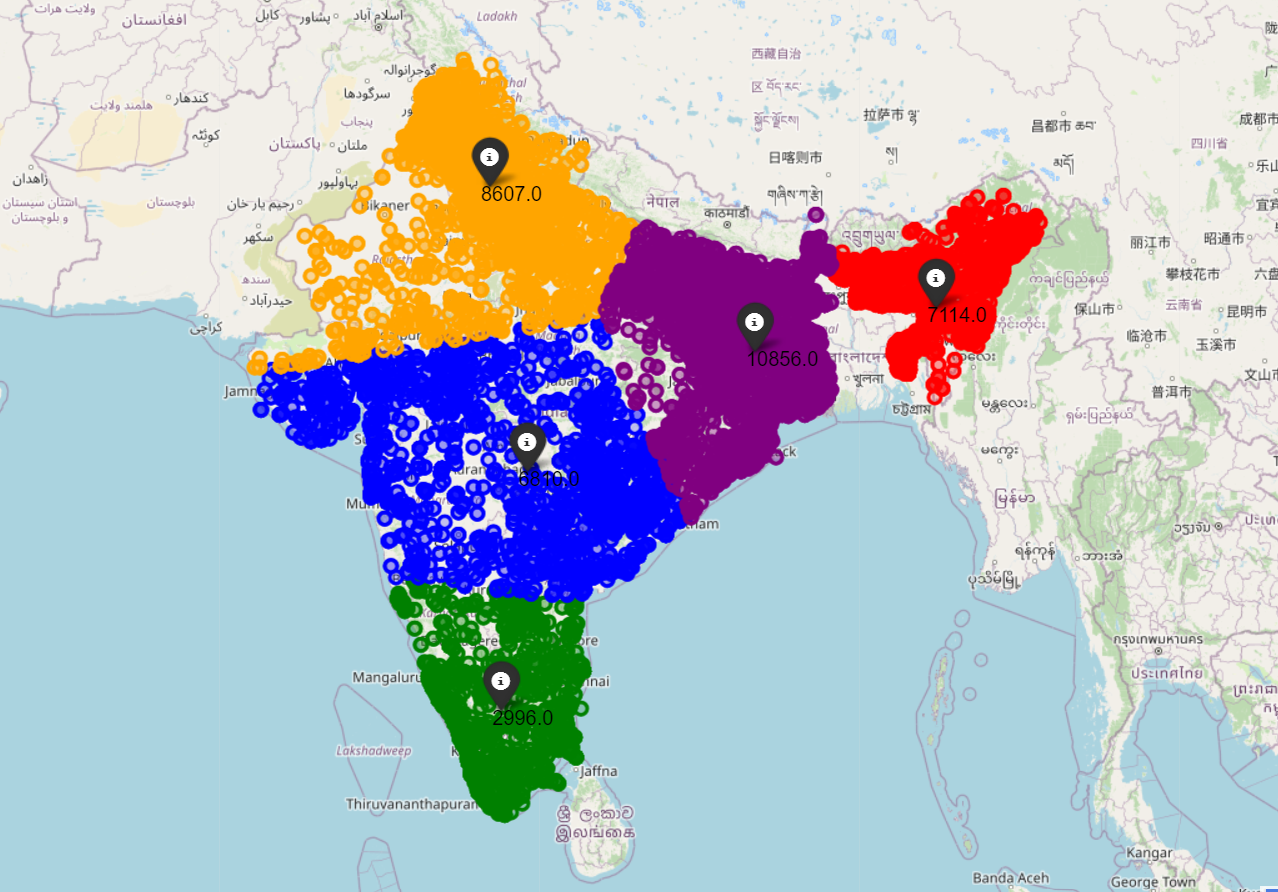
****

Fig 6. K-Means Clustering Revealing 5 Distinct clusters of Political Violence in India

The results of the cluster analysis show that **Cluster 2 (Purple) has the most incidents (10,856.0),**

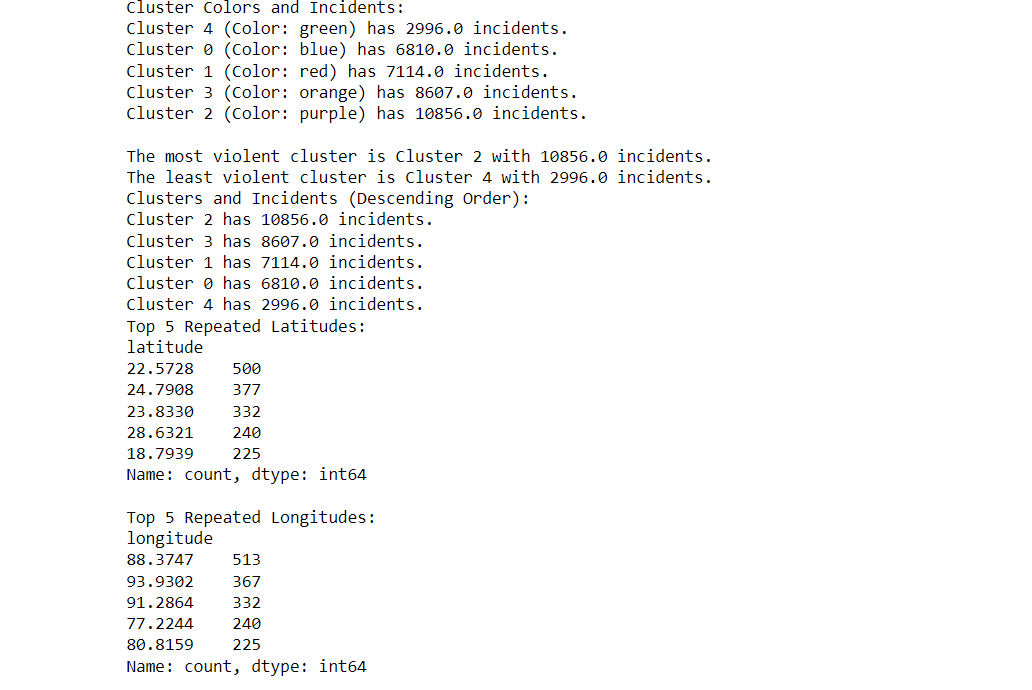


Fig 7. K-means clustering description for 5 clusters

**K-Means Clustering Results: 6 Clusters of Political Violence Incidents in India:**

The k-means clustering analysis has identified six distinct clusters of political violence incidents in India. Here's a breakdown of the information:

Cluster 0 (Yellow): This cluster has the most incidents with 10,397.

Cluster 2 (Purple): This cluster has 8164 incidents.

Cluster 3 (Orange): This cluster has 7059 incidents.

Cluster 5 (Red): This cluster has 4800 incidents.

Cluster 4 (Blue): This cluster has 3042 incidents.

Cluster 1 (Green): This cluster has the least incidents with 2921.

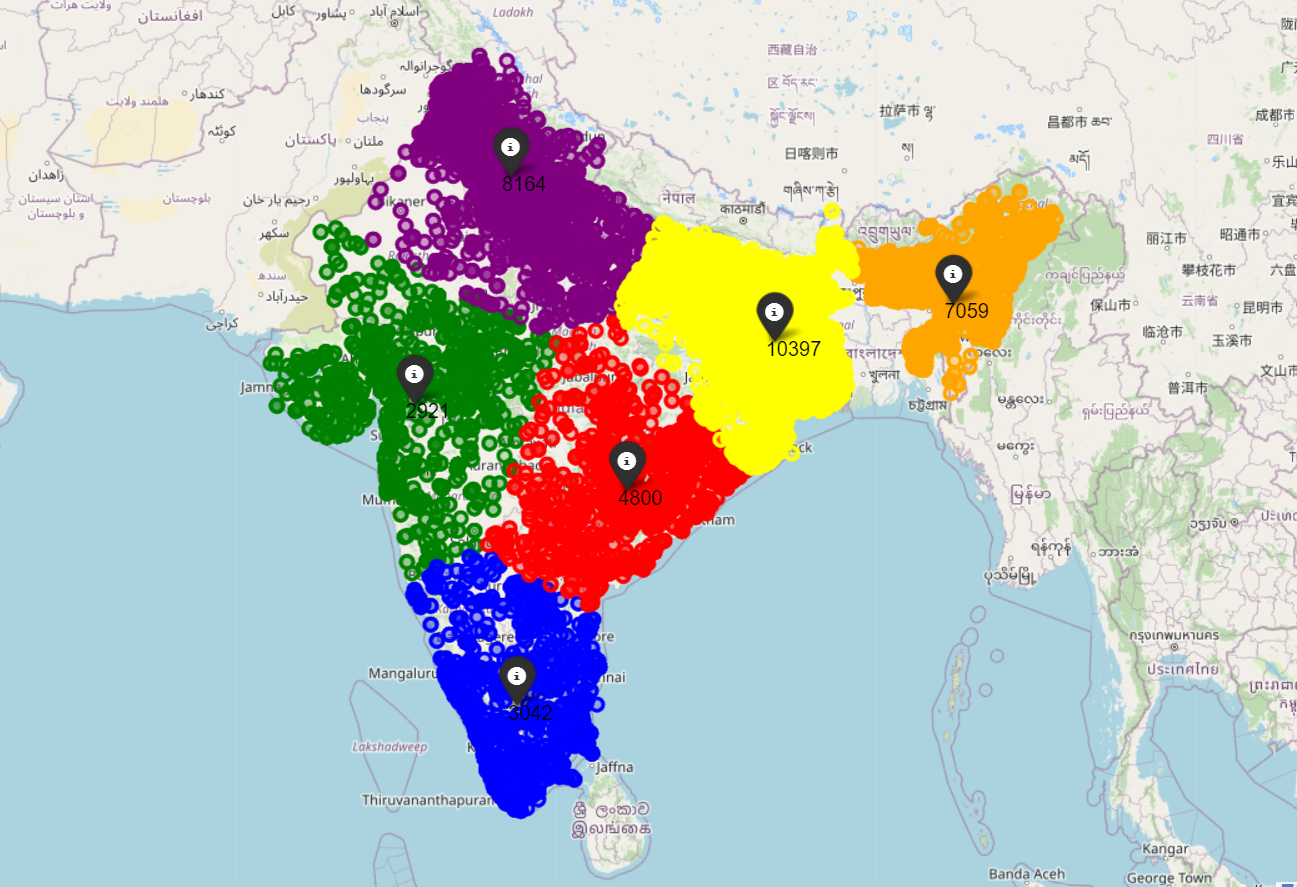
****

Fig 8. K-Means Clustering Results: 6 Clusters of Political Violence Incidents in India

### 

Fig 9. K-means clustering description for 6 clusters

**Identifying Top 5 Hotspots of Political Violence in India using K-Means Clustering**

The cluster analysis of political violence incidents in India revealed the top 5 hotspots of political violence:

* Kolkata, West Bengal (22.5728° N, 88.3747° E)
* Aizawl, Mizoram (24.7908° N, 93.9302° E)
* Silchar, Assam (23.833° N, 91.2864° E)
* New Delhi, Delhi (28.6321° N, 77.2244° E)
* Warangal, Telangana (18.7939° N, 80.8159° E)

These locations represent areas with a high concentration of political violence incidents, as identified through the cluster analysis. Further investigation and targeted interventions may be warranted in these regions to address underlying issues and mitigate the impact of political unrest.

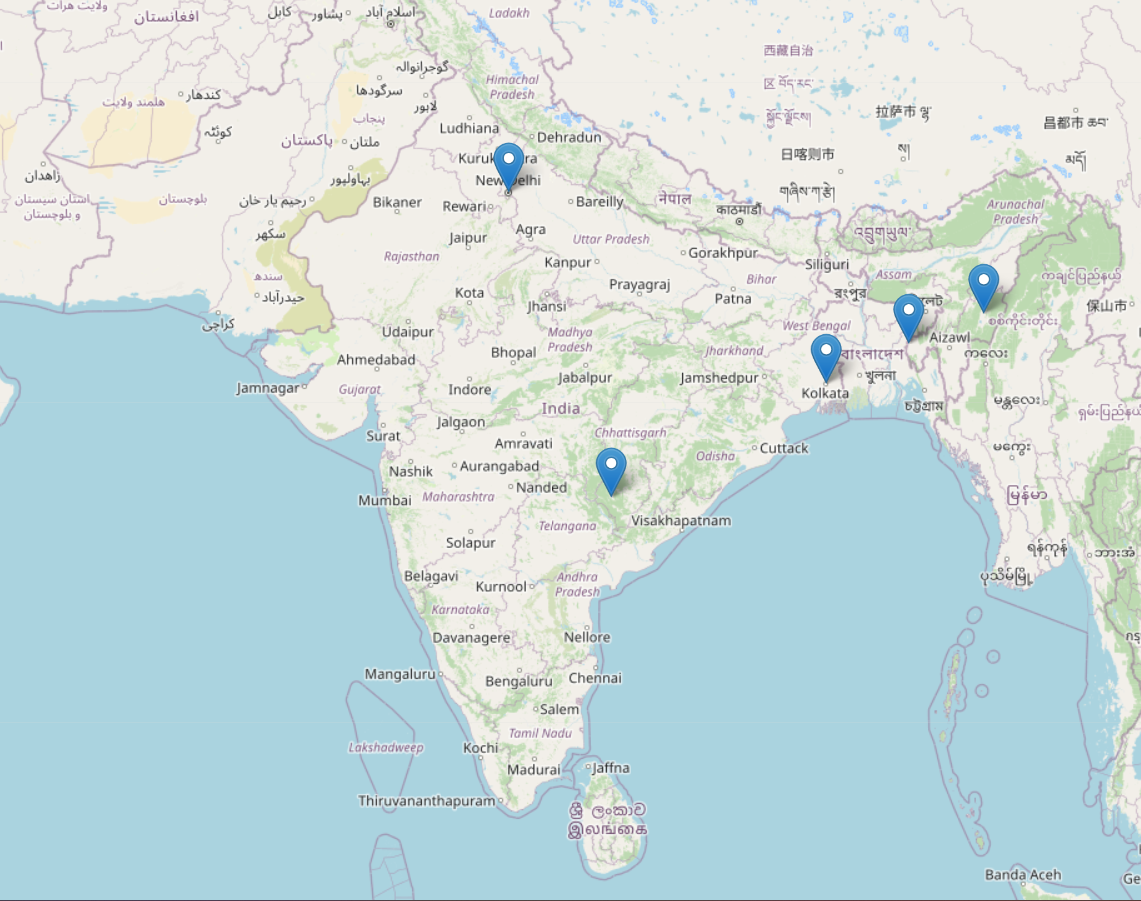
****

Fig 10. Top 5 Hotspots of Political Violence in India using K-Means Clustering

**Analysis of Event Types and Number of Fatalities through Cluster Analysis: -**

After conducting cluster analysis to investigate the correlation between event types, number of fatalities, and geographical locations of violent incidents, we identified distinct clusters. Here are the top 5 events causing fatalities in each of the 5 clusters:

**Cluster 1: Violence against civilians**

* Highest number of fatalities: 364

**Cluster 2: Battles**

* Highest number of fatalities: 2732

**Cluster 3: Violence against civilians**

* Highest number of fatalities: 1046

**Cluster 4: Violence against civilians**

* Highest number of fatalities: 1158

**Cluster 5: Riots**

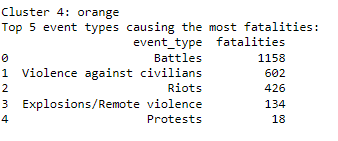
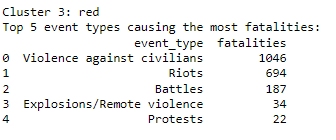
* Highest number of fatalities: 319

A white background with black text

Description automatically generated

A white background with black text

Description automatically generated



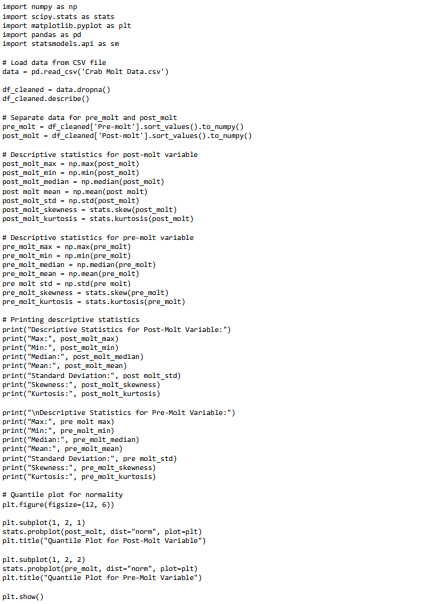
 A white background with black text

Description automatically generated



**Appendix C: DATA AND CODE**

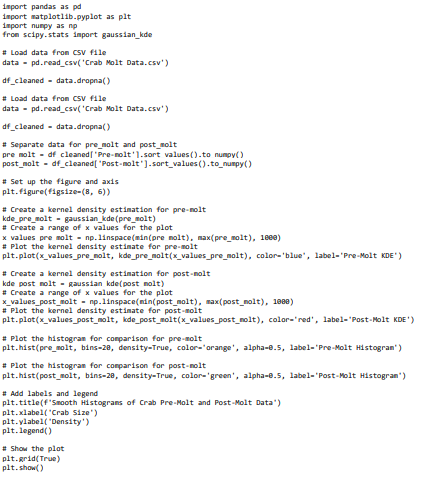
**Descriptive Statistics for Pre-Molt and Post-Molt Data :-**

****

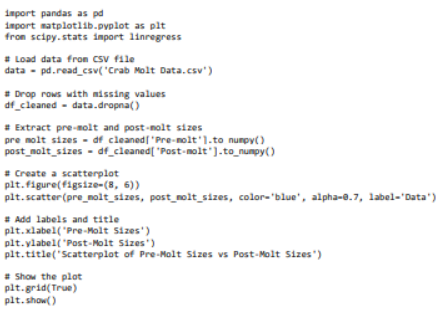
**Tests for Normality:**

****

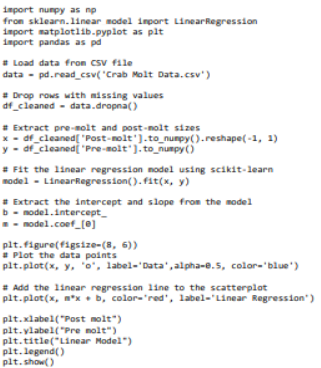
**Exploring Crab Sizes Before and After Molting: Histogram**

****

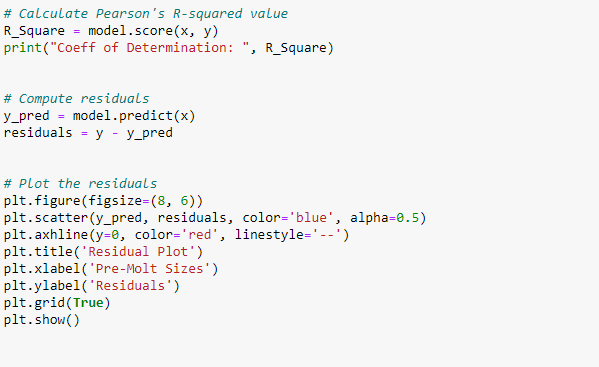
**Exploring Crab Sizes Before and After Molting: Scatter Plot**

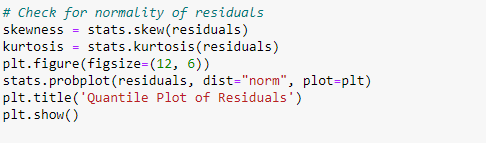
****

**Linear Model using Scatter Plot:**

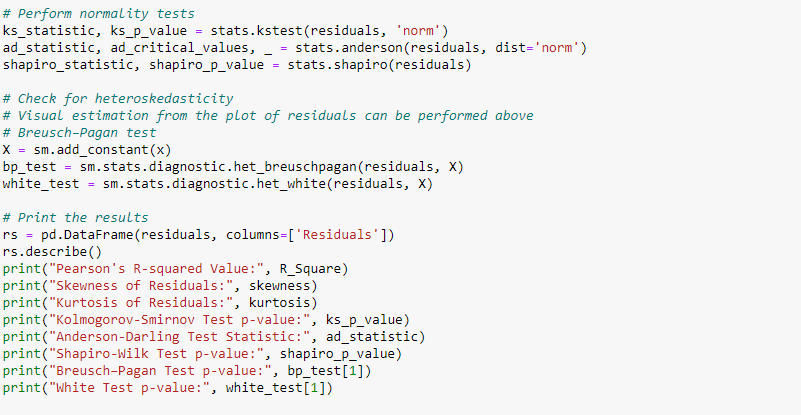
****

**Residual Plot:**

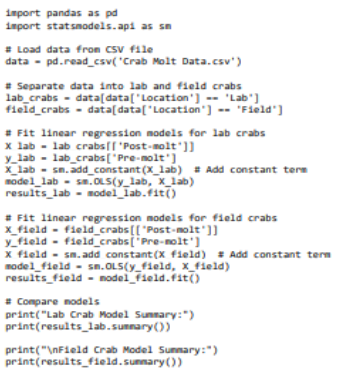
**Quantile-Quantile Plot:**

****

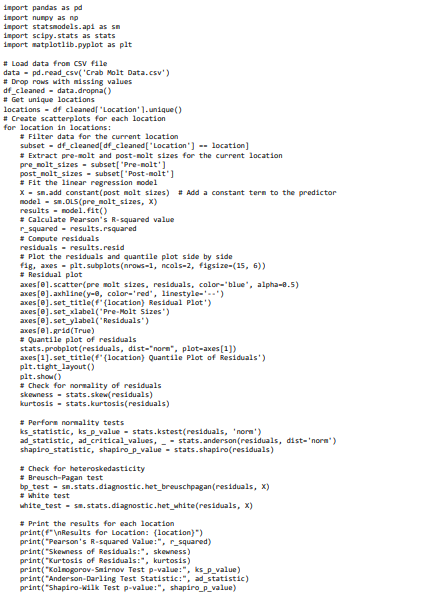
**Checks for Normality and Heteroskedasticity:**

****

**Analysis of Potentially Different Linear Models For Lab versus Field Crabs:**

****

**Checks for Normality and Heteroskedasticity For Lab Versus Field Crabs:**

****

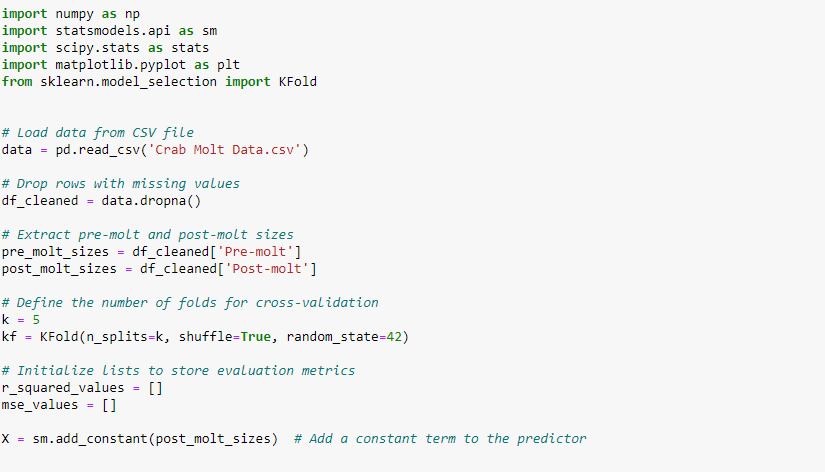
**Scatter Plot and Residual Plot For Lab Crabs:**

****

**Scatter Plot and Residual plot For Field Crabs:**

****

**K-Fold Cross Validation To Estimate Prediction Accuracy of Linear Models:**



**References:**

**[1]** *MTH 522 (Advanced Mathematical Statistics, sections 02)*

<https://mth522.wordpress.com/>

**[2]** <https://www.dropbox.com/scl/fi/rnhn88bfzbo26ugq408rj/IndiaPoliticalViolence-no-Islands_Kashmir_Ladakh.csv?rlkey=9676jgcbhjb2kv9azkg998jiv&e=1&dl=0>

**Contributions :**

Together, the four of us contributed equally and conducted a comprehensive study of the dataset.

**Supreeth Mohan:** Worked on the Issues, Discussion, Methods, Data Cleaning, Code, and Results sections. Also, used graphs to analyze the data using the various methods discussed in the report.

**Trina Xavier:** Worked on identifying issues, writing code for the analysis models, and producing the graphs.

**Aryan Bhalla:** Worked on the Issues, Findings, and Result sections. Plotted graphs and used various models and tests to analyze the data as discussed.

**Roshni Pal:** Worked on initial analyses, analyzing and looking for different models to fit the data on, testing various fits for their errors…