**Earthquake Data Analysis and Cluster Identification Report**

**1. Introduction**

This report details the analysis of earthquake data aimed at visualizing earthquake locations and identifying potential clusters within the data. The analysis utilizes various data processing techniques, interactive visualizations, and clustering algorithms to gain insights into earthquake patterns and distributions.

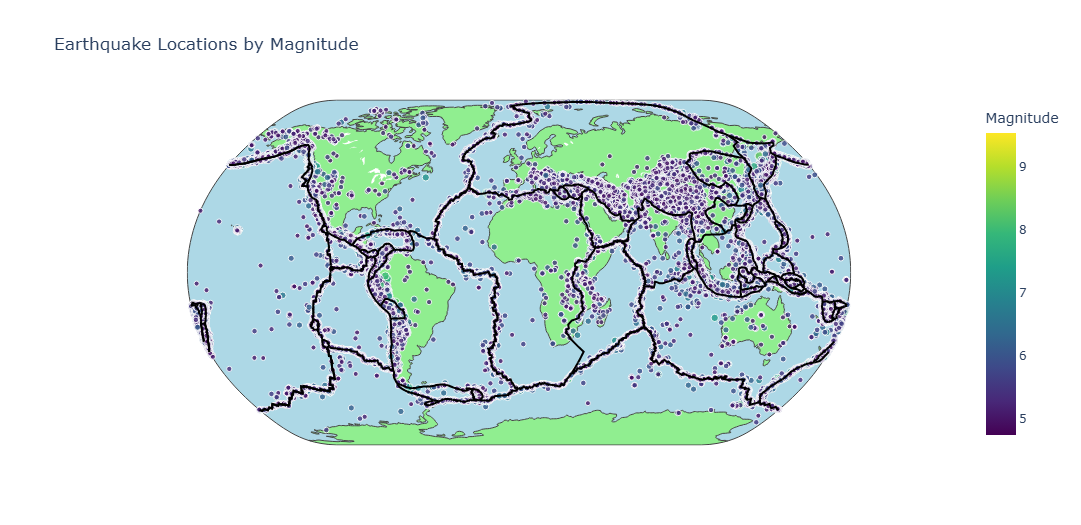
**2. Data and Methods**

**2.1 Data Loading and Preprocessing**

* Earthquake data is loaded from a CSV file (**DATA\_1.csv**).
* Data cleaning involves:
  + Splitting a single column into multiple columns.
  + Removing whitespace.
  + Renaming columns.
  + Converting latitude and longitude to numeric types.
* Plate boundary data is loaded from a JSON file (**PB2002\_boundaries.json**).

**2.2 Visualization**

* Interactive maps are created using the **plotly** library to visualize earthquake locations and plate boundaries.
* Earthquake magnitudes are represented by the **size and color** of the markers.



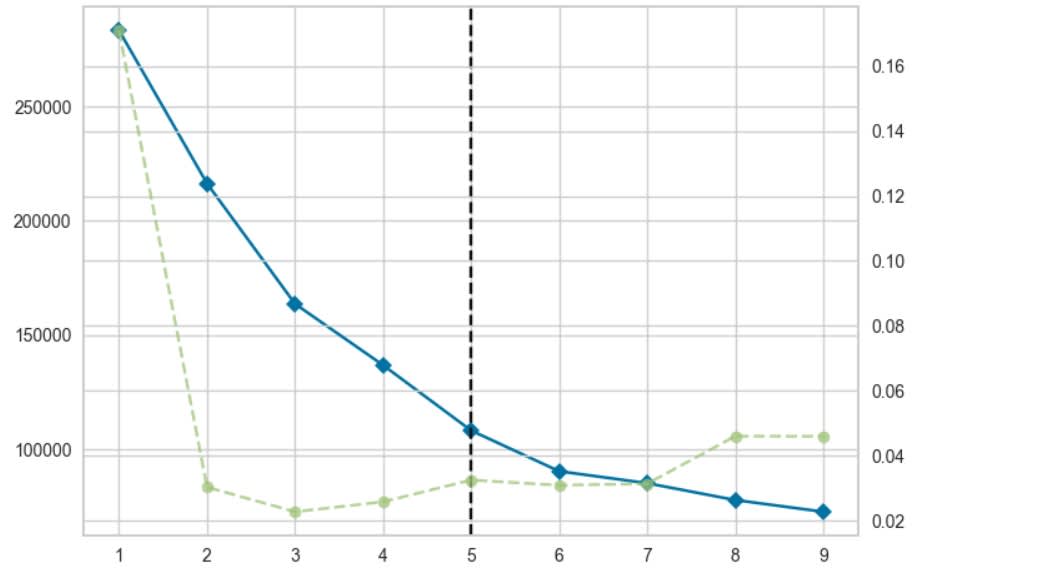
**2.3 Clustering Algorithms**

* **Feature Scaling:** Latitude, longitude, magnitude, and depth are scaled using **StandardScaler** before clustering.
* **K-Means Clustering:**
  + The **Elbow Method** is used to determine the optimal number of clusters.
  + K-Means clustering is then applied to the data.
* **DBSCAN Clustering:**
  + DBSCAN identifies clusters based on **density** and does not require specifying the number of clusters beforehand.
* **Gaussian Mixture Model (GMM):**
  + A **Gaussian Mixture Model** is used for clustering, assuming data points are generated from a mixture of Gaussian distributions.

**3. Results and Discussion**

**3.1 K-Means Clustering**

* The **Elbow Method** was used to determine the optimal number of clusters for the K-Means algorithm.
* **Figure 1: Elbow Plot for K-Means Clustering**
  + The elbow plot shows the relationship between the number of clusters and the distortion score.
  + The optimal number of clusters is selected at the **"elbow"** point of the curve.

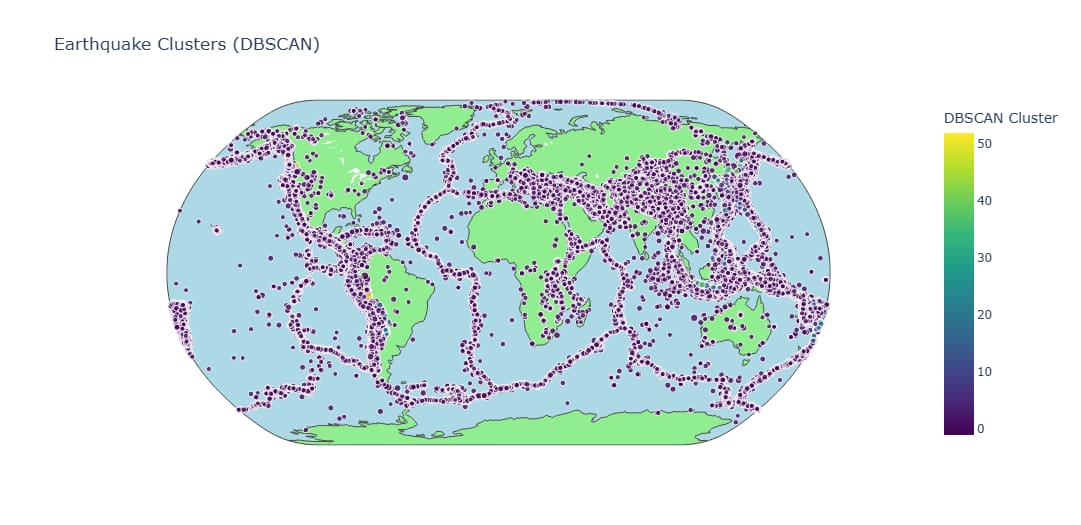


* **Figure 2: Earthquake Clusters after K-Means Clustering**
  + This figure visualizes the earthquake clusters identified by the **K-Means** algorithm on a geographical map.
  + Each cluster is represented by a different color.



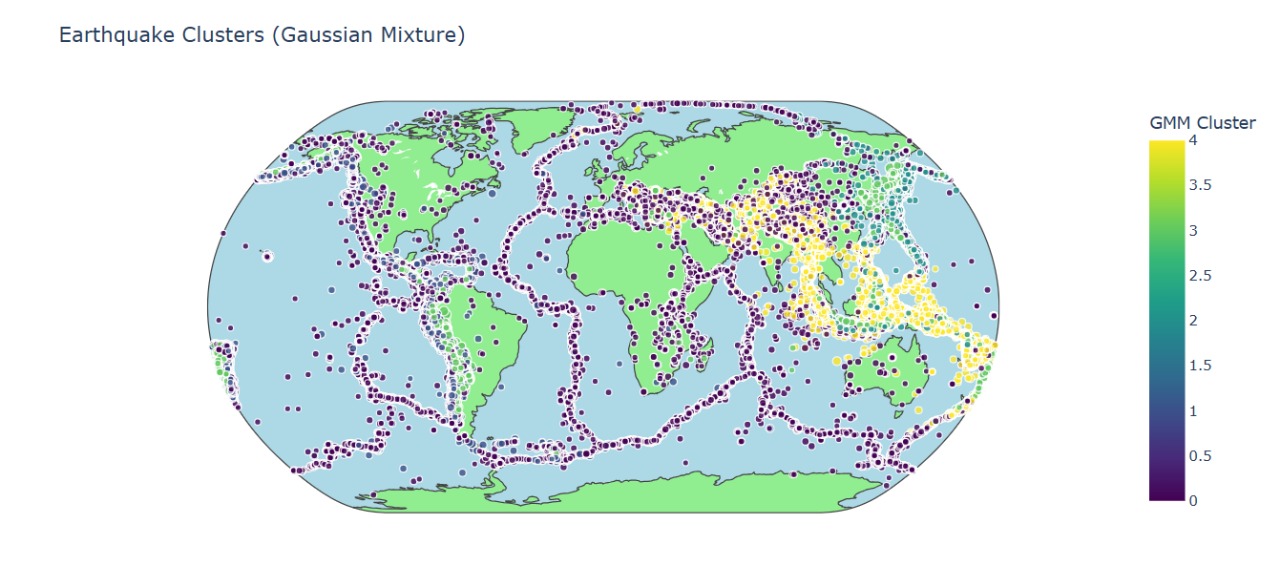
**3.2 DBSCAN Clustering**

* **Figure 3: Earthquake Clusters after DBSCAN Clustering**
  + This figure illustrates the clusters found by **DBSCAN**.
  + DBSCAN identifies clusters based on density without requiring a pre-defined number of clusters.
  + The color of each marker corresponds to the cluster assignment.



**3.3 Gaussian Mixture Model (GMM) Clustering**

* **Figure 4: Earthquake Clusters after Gaussian Mixture Model Clustering**
  + This figure displays the earthquake clusters identified using the **Gaussian Mixture Model**.
  + The GMM assumes that the data is generated from a mixture of Gaussian distributions.



**3.4 Evaluation Criteria**

**Evaluation Criteria for Clustering Methods**

When evaluating clustering performance, we consider multiple criteria to assess the quality of cluster formation. The primary metrics used in this analysis are **Silhouette Score** and **Davies-Bouldin Index**, which measure the compactness and separation of clusters. Below is a detailed explanation of these criteria:

**3.4.1. Silhouette Score**

* **Definition:** Measures how similar a point is to its assigned cluster compared to other clusters.
* **Range:** [-1, 1]
  + **High values (~1):** Points are well-clustered and distinct from other clusters.
  + **Close to 0:** Points are on the decision boundary between clusters.
  + **Negative values (~-1):** Points may be assigned to the wrong cluster.
* **Interpretation:** A higher silhouette score indicates well-defined, compact clusters with good separation.

**3.4.2. Davies-Bouldin Index (DBI)**

* **Definition:** Measures the average similarity ratio of each cluster with its most similar cluster.
* **Range:** [0, ∞] (Lower is better)
  + **Lower values (~0):** Better clustering with distinct and well-separated groups.
  + **Higher values:** Overlapping clusters with poor separation.
* **Interpretation:** A lower DBI value suggests that clusters are more compact and have greater separation from one another.

**How These Metrics Were Used**

The clustering evaluation was performed using three different methods:

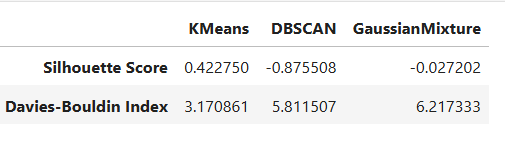
1. **KMeans:** A centroid-based clustering approach that partitions data into predefined clusters.
2. **DBSCAN:** A density-based clustering algorithm that groups data points based on their proximity.
3. **Gaussian Mixture Model (GMM):** A probabilistic clustering model that assumes data is generated from multiple Gaussian distributions.

Each method was applied to the dataset, and clusters were evaluated based on their **Silhouette Score** and **Davies-Bouldin Index**. Only clustering methods that produced at least two valid clusters were considered.

Based on the given evaluation scores, **KMeans** appears to be the best clustering method for this dataset.

**Explanation:**

1. **Silhouette Score**:
   * KMeans: **0.422750** (Higher is better)
   * DBSCAN: **-0.875508** (Negative indicates poor clustering)
   * GaussianMixture: **-0.027202** (Close to zero, indicating weak clustering)
   * **KMeans** has the **highest silhouette score**, meaning its clusters are more well-defined.
2. **Davies-Bouldin Index**:
   * KMeans: **3.170861** (Lower is better)
   * DBSCAN: **5.811507** (Higher indicates poor clustering)
   * GaussianMixture: **6.217333** (Highest, indicating worst performance)
   * **KMeans** has the **lowest Davies-Bouldin** Index, meaning it has better separation between clusters.

****

**4. Conclusion**

KMeans is the best clustering method for this dataset as it provides **better-defined clusters with more separation and compactness**, as indicated by the **highest Silhouette Score** and **lowest Davies-Bouldin Index**. DBSCAN and Gaussian Mixture performed poorly, likely due to unsuitable parameter selection or the nature of the dataset not aligning well with density-based or probabilistic clustering methods.

**TEAM DETAILS:**

HARSHA – CE22B007

S.A.V.SRIRAM – CE22B008

HARISH KUNDURU – CE22B009

J.V.S.SUDHEER – CE22B011

ROHITH MOKHARA – CE22B015

SATHWIK MERUGU – CE22B074

**THANKYOU**