# Clustering Analysis README

The README file provides an overview of the clustering analysis conducted on three datasets: dataset1, dataset2, and dataset3. The analysis includes K-Means clustering, evaluation metrics, and comparisons with shuffled data (dataset5). Below are details for each dataset and their respective results.

## Dataset 1

- \*\*Features\*\*: Annual Income ($) and Spending Score (1-100).

- \*\*Clustering Algorithm\*\*: K-Means with 3 clusters.

- \*\*Evaluation Metrics\*\*:

- Sum of Squared Errors (SSE).

- Silhouette Score.

- Davies-Bouldin Index.

- Dunn Index.

- \*\*Comparison with Shuffled Data (Dataset5)\*\*:

- Result: Clustering results for dataset1 are significantly better than random data (dataset5).

## Dataset 2

- \*\*Features\*\*: Teaching and Total Score.

- \*\*Clustering Algorithm\*\*: K-Means with 3 clusters.

- \*\*Evaluation Metrics\*\*:

- Sum of Squared Errors (SSE).

- Silhouette Score.

- Davies-Bouldin Index.

- Dunn Index.

- \*\*Comparison with Shuffled Data (Dataset5)\*\*:

- Result: Clustering results for dataset2 are significantly better than random data (dataset5).

## Dataset 3

- \*\*Features\*\*: Latitude and Longitude.

- \*\*Clustering Algorithm\*\*: K-Means with 3 clusters.

- \*\*Evaluation Metrics\*\*:

- Sum of Squared Errors (SSE).

- Silhouette Score.

- Davies-Bouldin Index.

- Dunn Index.

- \*\*Comparison with Shuffled Data (Dataset5)\*\*:

- Result: Clustering results for dataset3 are not significantly better than random data (dataset5).

## Conclusion

The clustering analysis was conducted on three distinct datasets: dataset1, dataset2, and dataset3, each representing different types of data. The primary goal was to identify meaningful clusters within these datasets and evaluate the quality of clustering results. The analysis involved the application of the K-Means clustering algorithm with three clusters for each dataset.

- Dataset1 and dataset2 demonstrate successful clustering with clear separation from random data, indicating meaningful clusters.

- Dataset3 shows clustering results that are not significantly better than random data, suggesting potential areas for improvement or alternative clustering techniques.

### Dataset 1

Dataset1 consists of information related to annual income and spending score, making it ideal for customer segmentation or market targeting. The K-Means clustering algorithm successfully identified three distinct clusters within this dataset. The evaluation metrics, including Sum of Squared Errors (SSE), Silhouette Score, Davies-Bouldin Index, and Dunn Index, indicated that the clustering results for dataset1 are not only meaningful but also significantly better than random data (dataset5). This suggests that the chosen features and clustering parameters are well-suited for uncovering valuable insights within the data.

### Dataset 2

In the case of dataset2, which contains attributes related to teaching and total score, K-Means clustering was applied to identify potential clusters. The analysis demonstrated that the clustering results for dataset2 are also meaningful and significantly better than random data (dataset5). This indicates that the features chosen for clustering and the number of clusters were appropriate for capturing relevant patterns within the data. The successful clustering of dataset2 provides valuable insights for education or research-related applications.

### Dataset 3

Dataset3 contains geographical coordinates, specifically latitude and longitude information. Clustering geographic data is often challenging, and in this case, the results were not as clear-cut. The clustering analysis suggested that the results for dataset3 are not significantly better than random data (dataset5). This outcome indicates that additional considerations, such as alternative clustering techniques or feature engineering, may be necessary to extract meaningful clusters from geographic data. Despite this, the analysis provides a starting point for further exploration of location-based insights.

In summary, the clustering analysis revealed that datasets 1 and 2 yielded meaningful and significantly better results compared to random data. In contrast, dataset3, which contains geographic information, presented a more complex clustering challenge. The results serve as a foundation for understanding and interpreting the data and can guide further refinements in the clustering process or the selection of features. It's important to note that the effectiveness of clustering may vary depending on the nature of the data and the intended application. These findings provide valuable insights into the quality of clustering results and point towards potential areas for improvement.