Clustering Assignment

Objective:
Perform clustering analysis using K-Means and HDBSCAN on the provided dataset to identify the optimal number of clusters.

Steps to Follow:

1. Import Libraries:
   Gather all necessary libraries for data manipulation, visualization, and clustering.

2. Load and Understand the Dataset:
   Import the dataset and examine its structure, exploring the features and data types. Conduct exploratory data analysis (EDA) to gain insights into the dataset by visualizing distributions, checking correlations, and identifying patterns.

3. Data Preprocessing:
   - Handle Missing Data: Identify and appropriately address any missing values in the dataset.
   - Standardize the Data: Scale the features so that they have a mean of 0 and a standard deviation of 1 using appropriate scaling methods.
   - Normalize the Data (Optional): If needed, normalize the data to bring all features onto a similar scale.

4. K-Means Clustering:
   - Implement K-Means with K = 6 clusters, assigning each data point to one of these clusters.
   - Evaluate the optimal number of clusters using the elbow method and silhouette score, which helps in assessing the quality of the clusters.
   - Visualize the resulting clusters, including their centroids, to understand the distribution and separation of clusters.

5. HDBSCAN Clustering:
   - Implement HDBSCAN by experimenting with different parameter values, such as `min_cluster_size` and `min_samples`, to find the most suitable clustering configuration.
   - Utilize HDBSCAN's adaptive parameters to automatically determine optimal values like `epsilon`, adjusting as necessary based on the dataset's distribution.
   - Visualize the resulting clusters, including any noise points (data points that do not belong to any cluster).

6. Comparison and Evaluation:
   - Compare the results obtained from K-Means and HDBSCAN in terms of their effectiveness on the dataset. Consider factors like noise handling, the shape of clusters, and cluster density.
   - Discuss which method performs better, providing reasoning based on the dataset’s characteristics, such as how well each method handles noise, irregular cluster shapes, and overall clustering performance.

Dataset- https://www.kaggle.com/datasets/shrikantuppin/east-west-airlines

For HDBSCAN refer to this article-
https://www.geeksforgeeks.org/hdbscan/