Final Project Report - Big Data

GitHub Repository

https://github.com/DeYi0124/BDA-final

Algorithm Used

We used the KMeans clustering algorithm with a predefined number of clusters based on 4n-1. KMeans is suitable for large datasets with known cluster counts.

Suitability for the Dataset

Given the high dimensionality (4D and 6D) and the clearly defined cluster count, KMeans is effective and computationally efficient.

Handling High-Dimensional Data

All features were standardized using z-score normalization (StandardScaler) to ensure fair distance measurements in clustering.

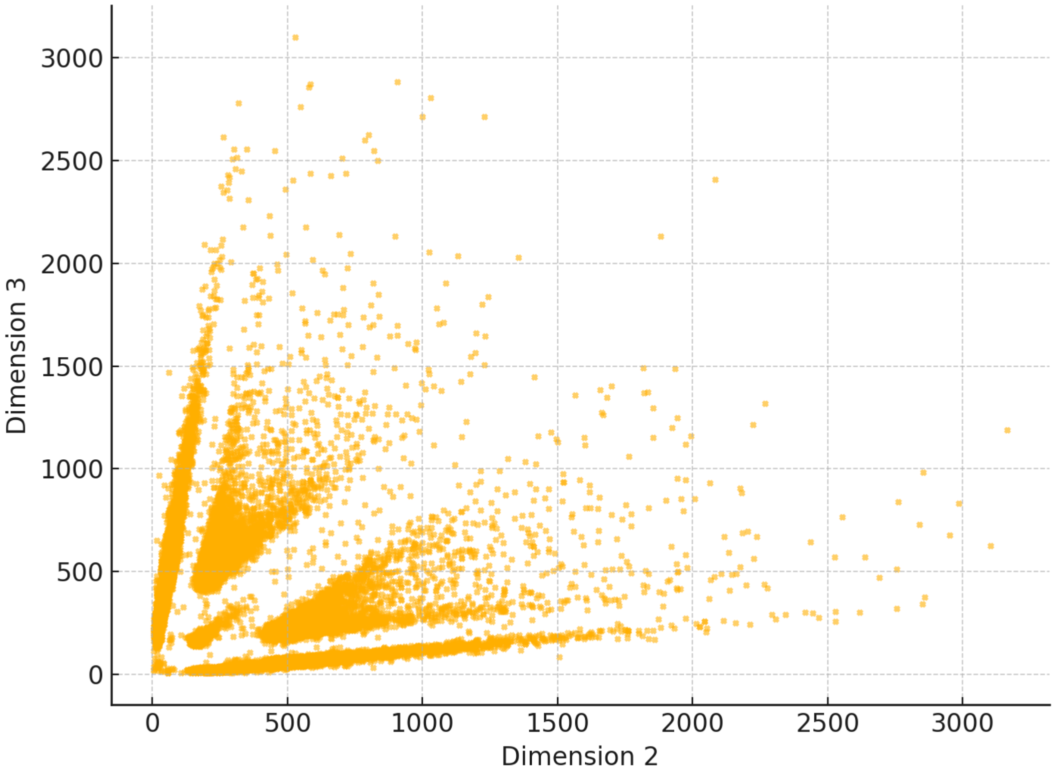
Preprocessing and Hyperparameters

Data was scaled using StandardScaler. We set the number of clusters as 15 (for 4D) and 23 (for 6D). KMeans used 10 restarts to avoid poor local minima.

Visual Analysis of Public Dataset

By visualizing Dimension 2 vs 3 of the public dataset, we observed that the data appears to form around five visible clusters, suggesting non-random spatial grouping.

Final Project Report - Big Data


Public Dataset: Dimension 2 vs Dimension 3