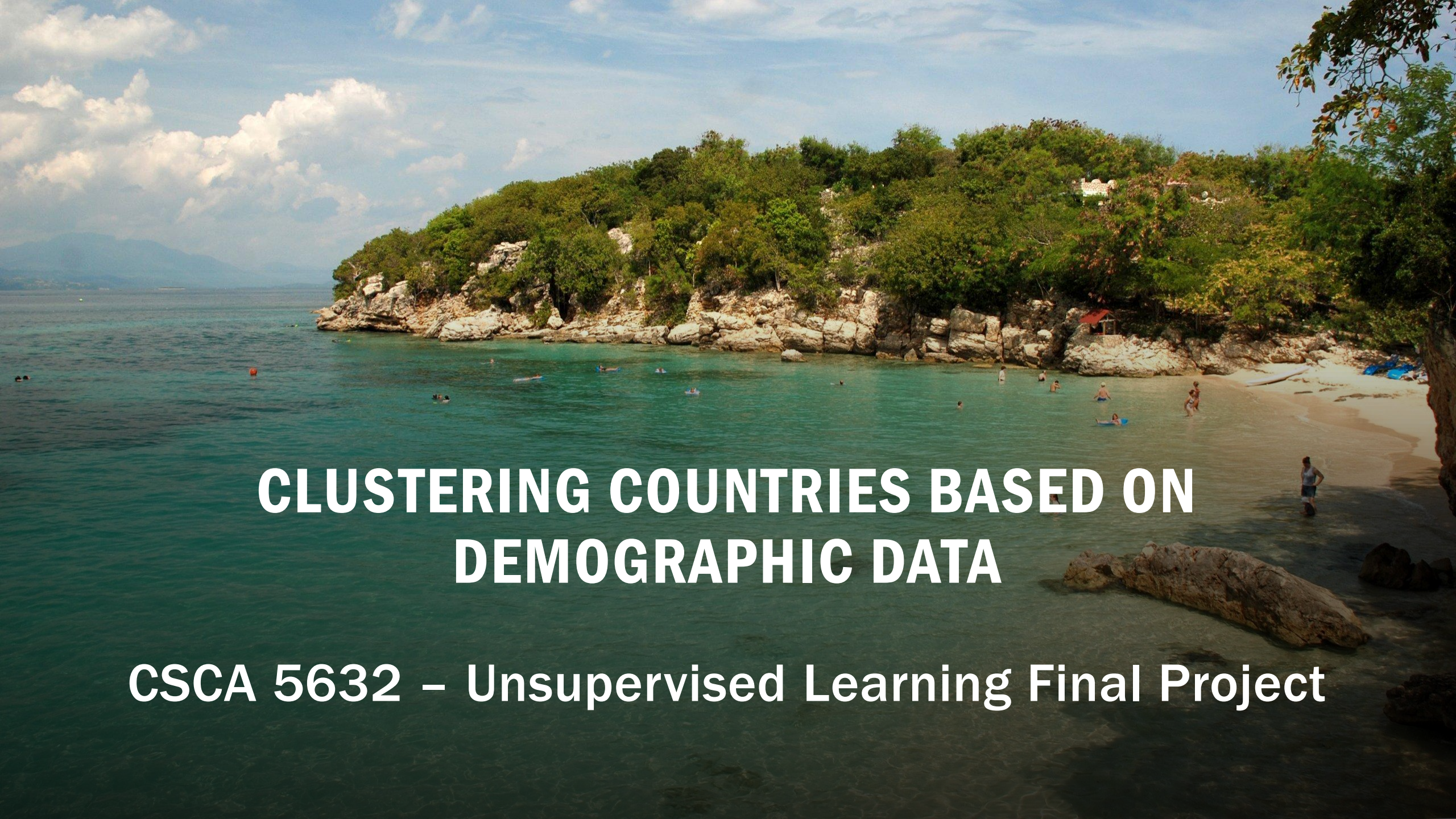


CLUSTERING COUNTRIES BASED ON DEMOGRAPHIC DATA

CSCA 5632 – Unsupervised Learning Final Project

PROBLEM STATEMENT AND OVERVIEW

Can unsupervised clustering techniques uncover meaningful segments among countries based on socio-economic and demographic indicators, and which algorithm produces the most coherent and interpretable groupings?

Project Goals:

- Explore and prepare the dataset.
- Apply different clustering techniques.
- Evaluate and compare the clustering results.

167 countries

10 variables properties, at a snapshot in time

Key features include:

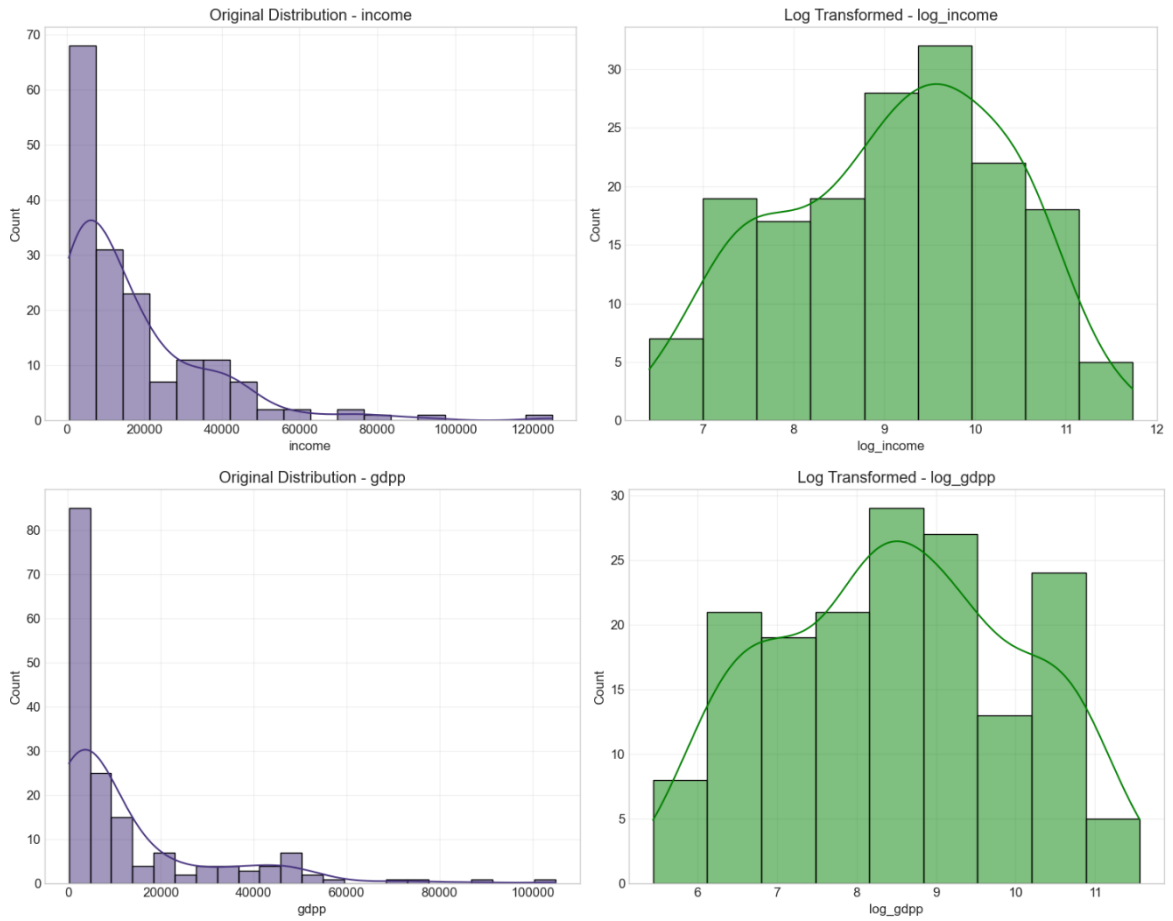
- Child mortality
- Imports + exports
- GDP per capita
- Inflation

DATASET

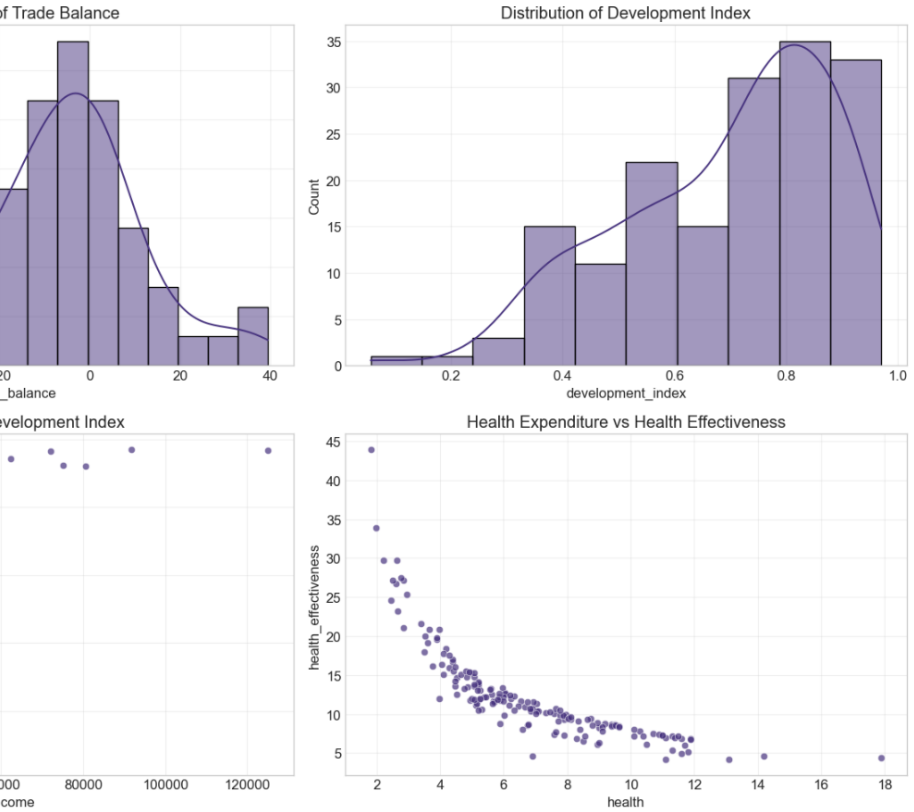
EXPLORATORY DATA ANALYSIS HIGHLIGHTS

Several non-normally distributed variables:

- **These were reviewed for normality and skew**
- **Most were normalized ahead of feature engineering**



FEATURE ENGINEERING



Created features to capture interrelationships and key variables:

- Trade balance (i.e. exports relative to imports)
- Healthcare expenditure vs. healthcare effectiveness
- Created a “development index” variable (similar to an HDI-like value)

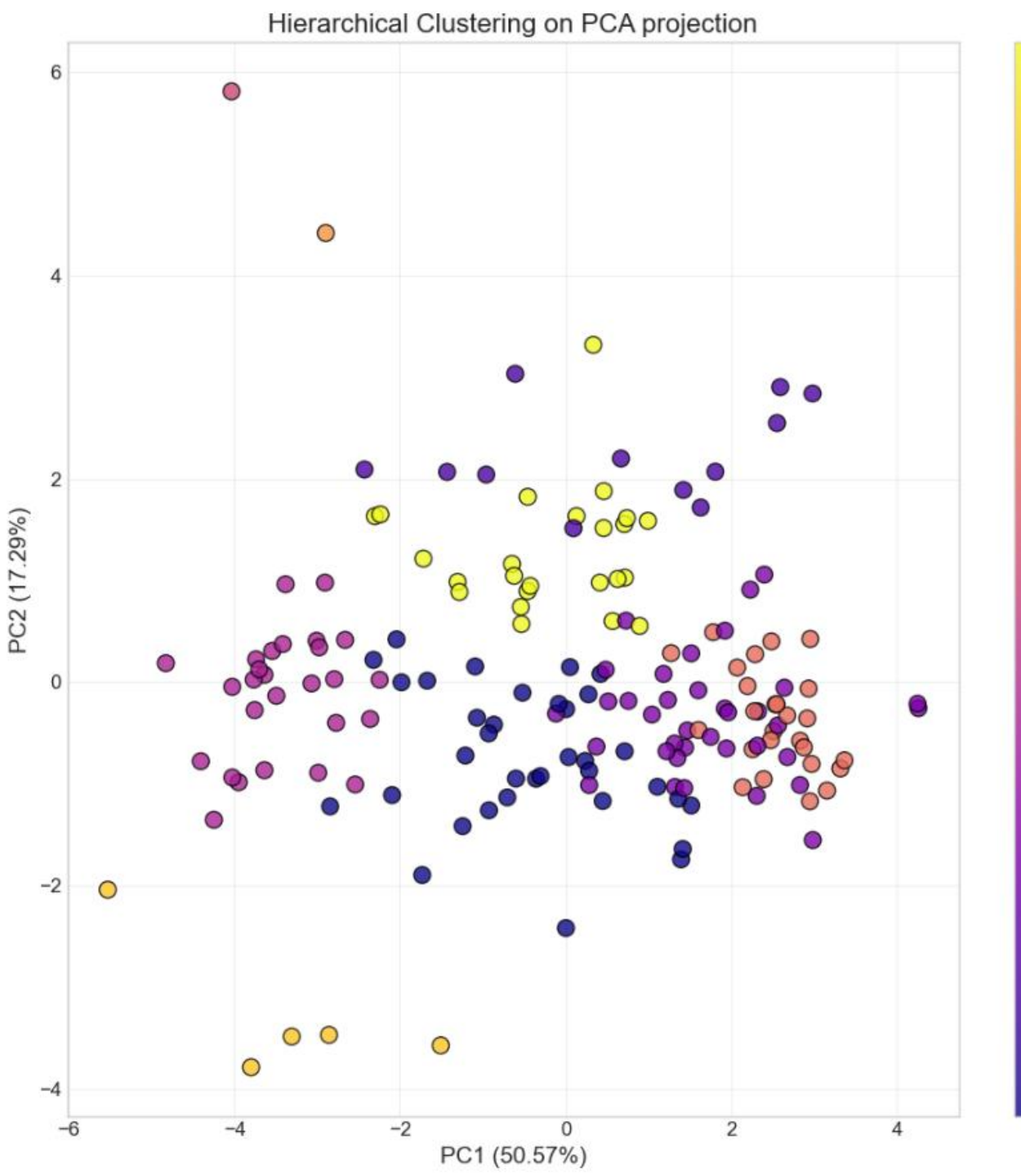
MODELLING APPROACH

Implemented and compared four models:

- K-means Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering)
- Gaussian Mixture Model (GMM)

Evaluation Criteria:

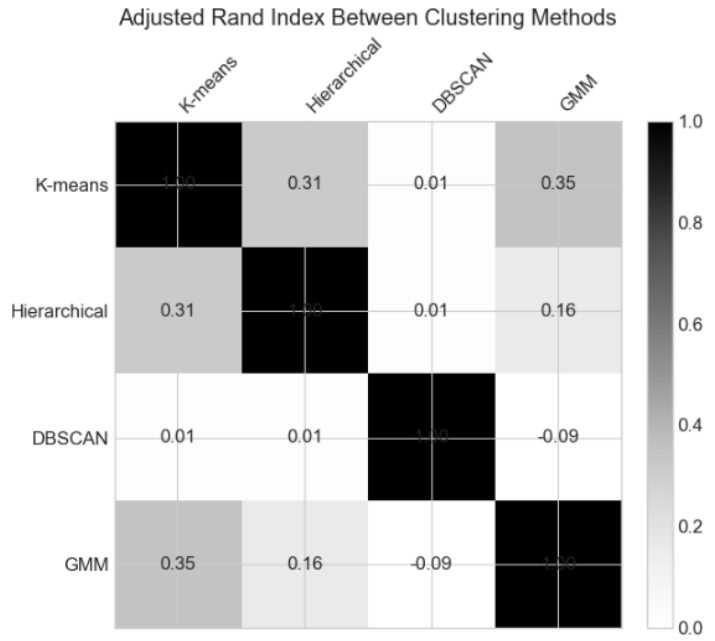
- Silhouette score.
- Inertia (within-cluster sum of squares).
- Visual analysis of cluster separation.



MODEL PERFORMANCE

Two measures reviewed:

- Adjusted Rand Index
- Silhouette Score comparison



CONCLUSIONS AND RECOMMENDATIONS

Gaussian Mixture Models remain the strongest performer once noise is included, thanks to their probabilistic cluster assignments.

However, also the fewest clusters given selected hyperparams – potentially low explanatory power

K-Means and Hierarchical hold up moderately well – there is regional overlap

DBSCAN's negative score highlights that the noise label truly represents outliers: these points do not form a cohesive cluster and degrade the overall silhouette when forced into one group.