A CLUSTERING-BASED APPROACH TO IDENTIFYING DEVELOPMENT PATTERNS FOR AID DECISION-MAKING

FINAL PROJECT — DTSA-5510
UNSUPERVISED ALGORITHMS IN MACHINE LEARNING

JINNAJATE ACHALAPONG

INTRODUCTION & PROBLEM DEFINITION

Introduction

HELP International, a global humanitarian NGO, has recently raised $10 million in funding to support developing nations. In order to maximize impact, it is crucial to identify countries with similar socio-economic and health conditions that are most in need of aid.

Problem Definition

This project applies unsupervised machine learning techniques
specifically clustering algorithms to group countries based on development indicators.
The goal is to uncover latent structures in the data to:

- Identify homogeneous groups of countries
- Support data-driven decisions on international aid allocation
- Enhance the effectiveness and fairness of resource distribution

INITIAL DATA INSPECTION — INSIGHTS

Exploratory Inspection

- 167 countries × 10 features
- 9 numerical, 1 categorical (country)
- No missing values
- Used head(), info(), describe() for overview

Key Insights

- Child Mortality: 2.6–90.2
- Life Expectancy: 32–83 years
- Income & GDP per Capita: Wide range
- Inflation: -4.2% to 104%
- Fertility Rate: Up to 7.5

```
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   country     167 non-null    object
 1   child_mort  167 non-null    float64
 2   exports     167 non-null    float64
 3   health      167 non-null    float64
 4   imports     167 non-null    float64
 5   income      167 non-null    int64
 6   inflation   167 non-null    float64
 7   life_expec  167 non-null    float64
 8   total_fer   167 non-null    float64
 9   gdpp        167 non-null    int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

Summary Statistics:

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00
mean	38.27	41.11	6.82	46.89	17144.69	7.78	70.56	2.95	12964.16
std	40.33	27.41	2.75	24.21	19278.07	10.57	8.89	1.51	18328.70
min	2.60	0.11	1.81	0.07	609.00	-4.21	32.10	1.15	231.00
25%	8.25	23.80	4.92	30.20	3355.00	1.81	65.30	1.79	1330.00
50%	19.30	35.00	6.32	43.30	9960.00	5.39	73.10	2.41	4660.00
75%	62.10	51.35	8.60	58.75	22800.00	10.75	76.80	3.88	14050.00
max	208.00	200.00	17.90	174.00	125000.00	104.00	82.80	7.49	105000.00

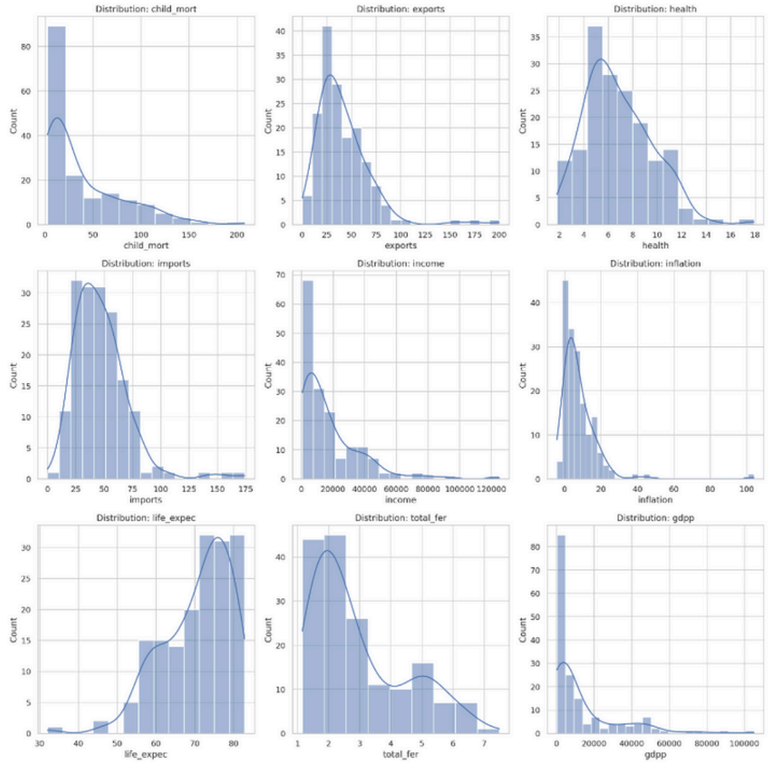
SUMMARY OF EXPLORATORY DATA ANALYSIS (EDA)

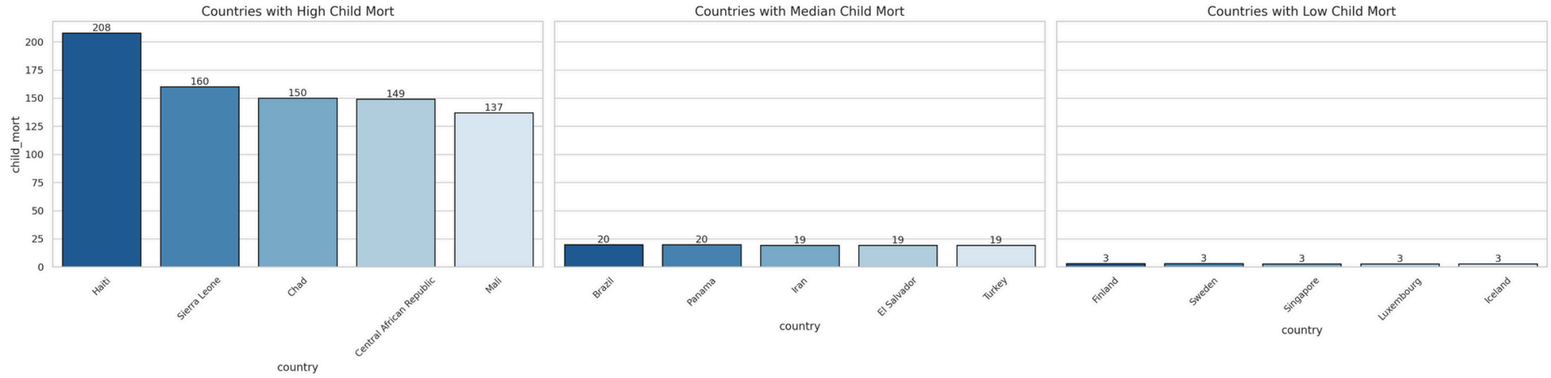
🌍 Global Development Patterns

- African countries show consistently poor outcomes:
- → High child mortality, low life expectancy, high fertility, high inflation
- → Suggests strong case for targeted aid
- Asian & European countries generally perform better on health and demographic metrics
- Haiti has highest child mortality despite being outside Africa

💰 Economy, Health & Trade

- High health spending ≠ better outcomes (e.g., USA vs life expectancy leaders)
- Top income & GDPP: Qatar, Luxembourg, Switzerland
- Trade-reliant nations (e.g., Singapore) lead in both exports and imports
- African countries dominate lowest ranks in income, GDPP, and inflation stability





FEATURE ENGINEERING & DIMENSIONALITY REDUCTION

Feature Grouping

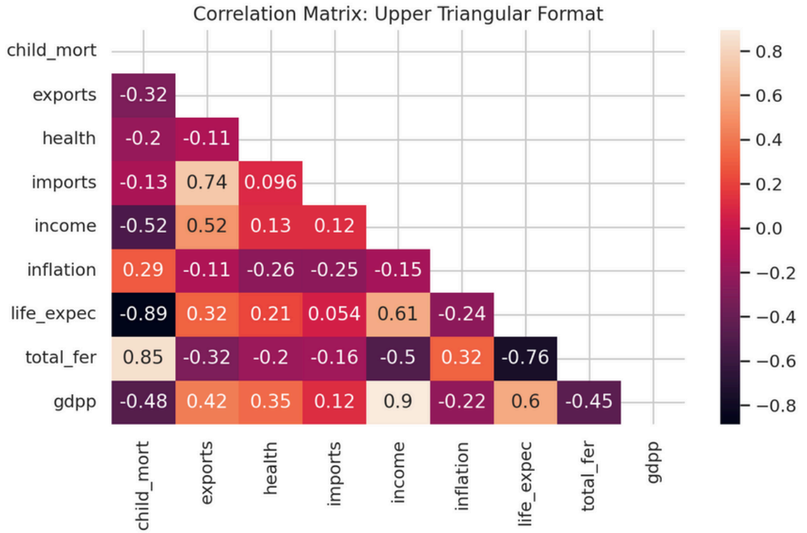
- Constructed 3 composite features to enhance interpretability:
- Health: child_mort, health, life_expec, total_fer
- Trade: exports, imports
- Finance: income, inflation, gdpp
- → All components normalized by mean before summing

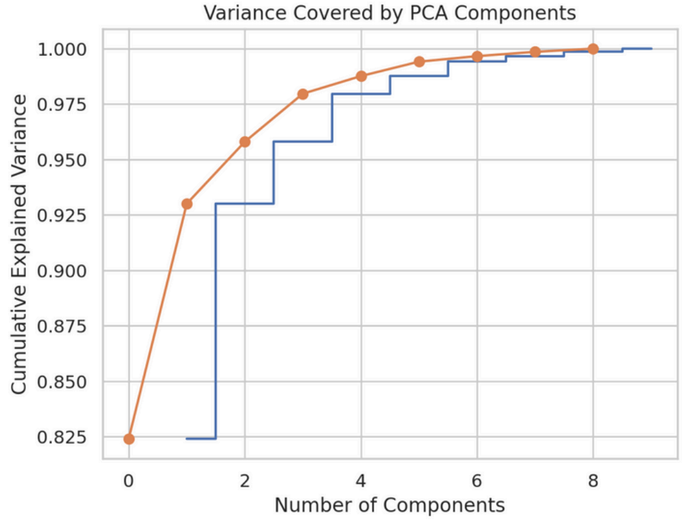
Data Scaling

- Applied Min-Max Scaling to all features (0–1 range)
- Ensures fair treatment in distance-based clustering
- Standardization used for Gaussian-like variables (e.g., health for PCA)

Principal Component Analysis (PCA)

- PCA performed on normalized features (excluding country)
- First 3 components capture most of the variance
- → Used for dimensionality reduction before clustering

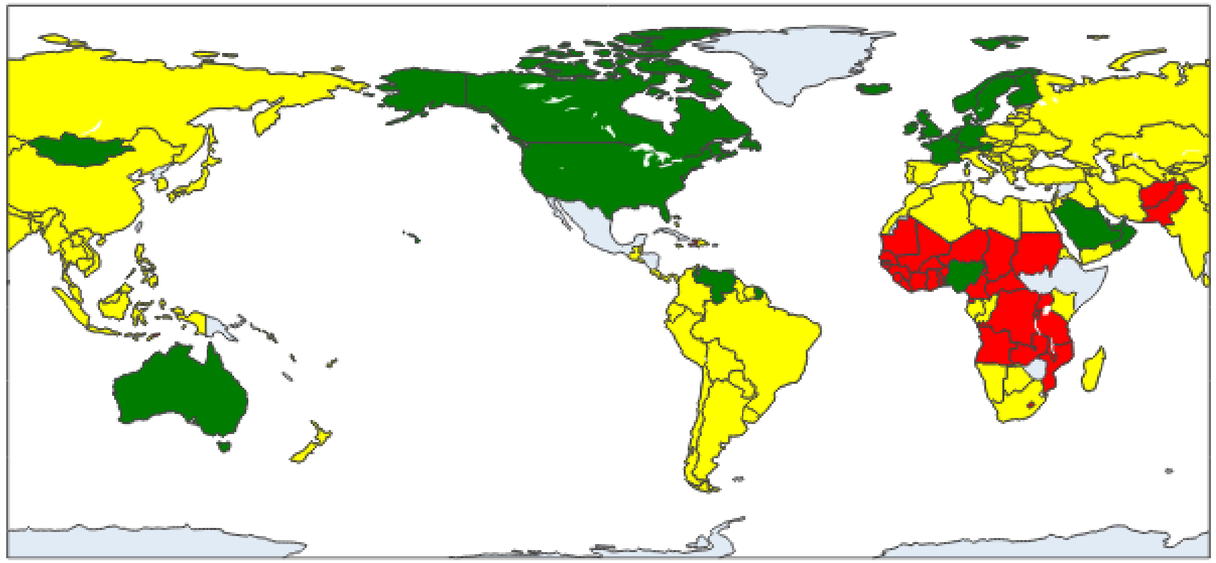


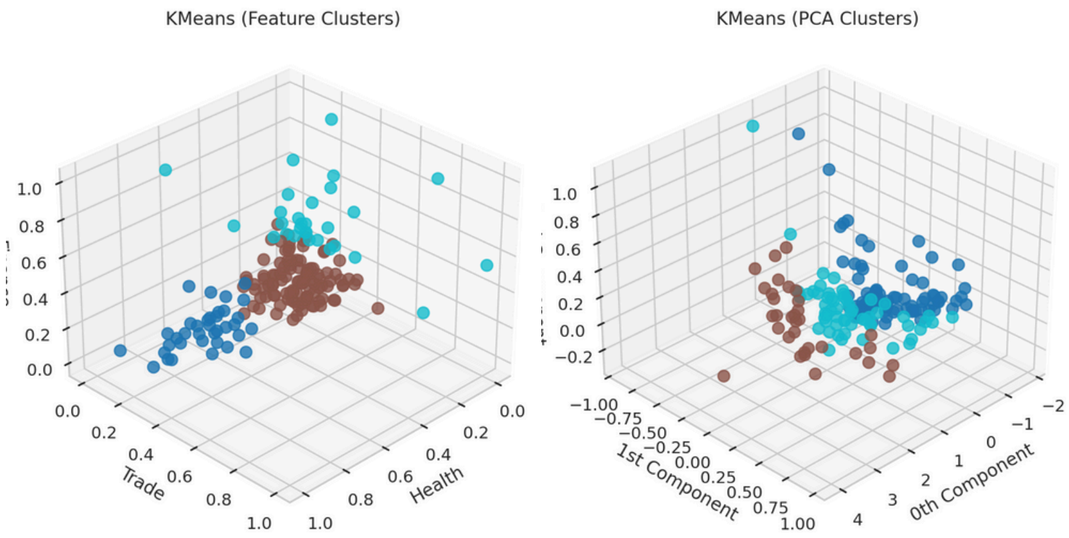


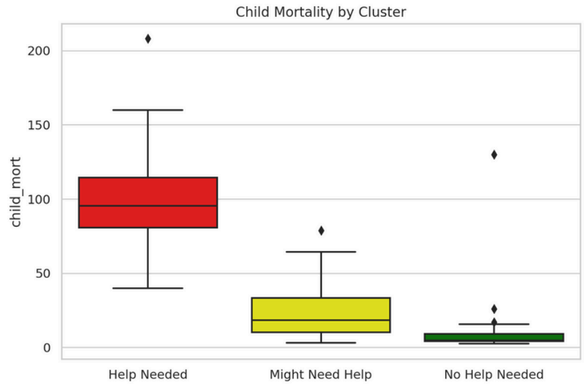
K-MEANS CLUSTERING RESULTS

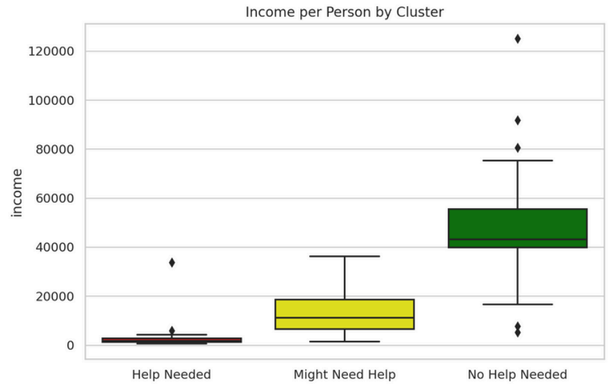
Quantitative Evaluation

Metric	Feature-based	PCA-based	Better?
Silhouette Score	**0.452**	0.392	Feature-based
Calinski–Harabasz Score	125.6	**202.5**	PCA-based
Davies–Bouldin Index	0.888	**0.856**	PCA-based
Inertia (SSE)	6.45	56.60	Not directly comparable[1]





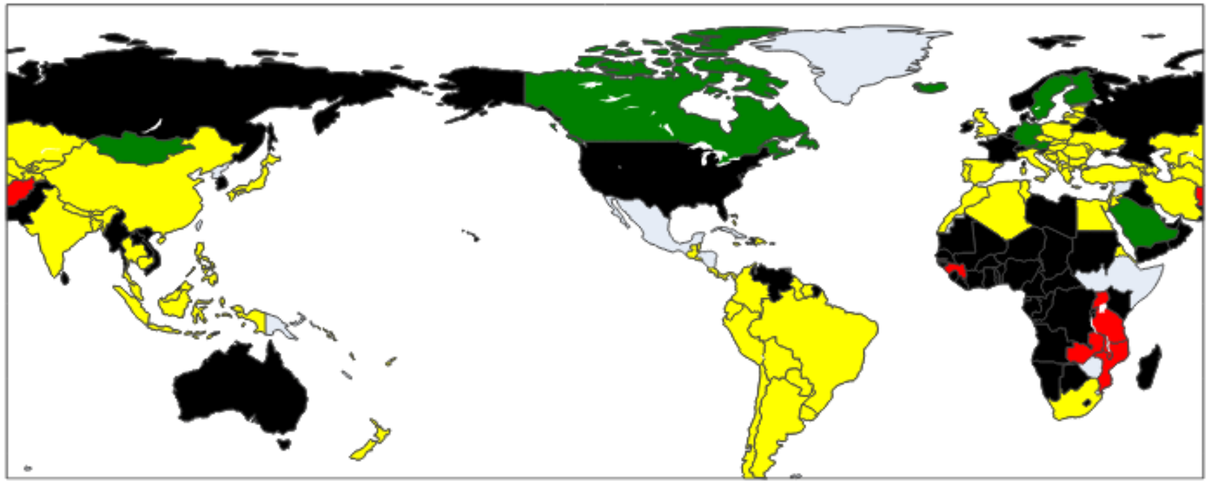


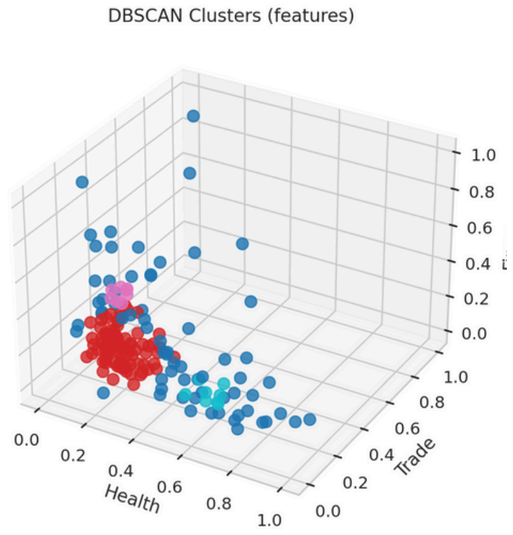


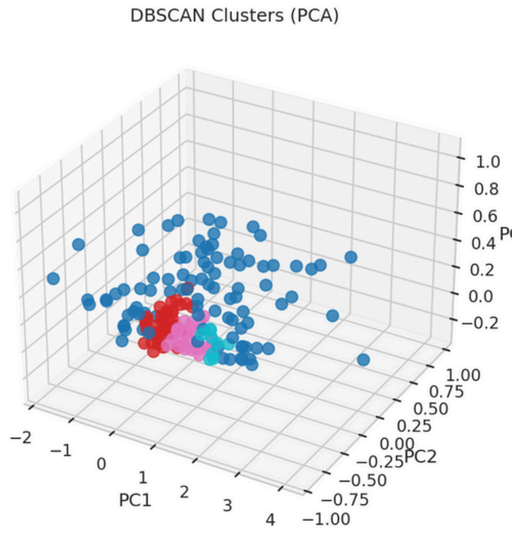
DBSCAN CLUSTERING RESULTS

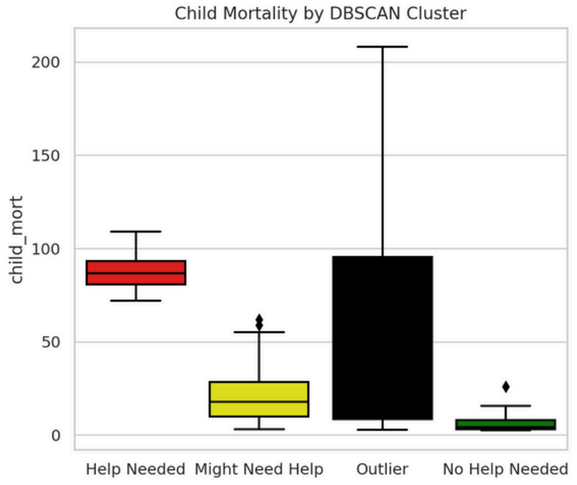
Quantitative Evaluation

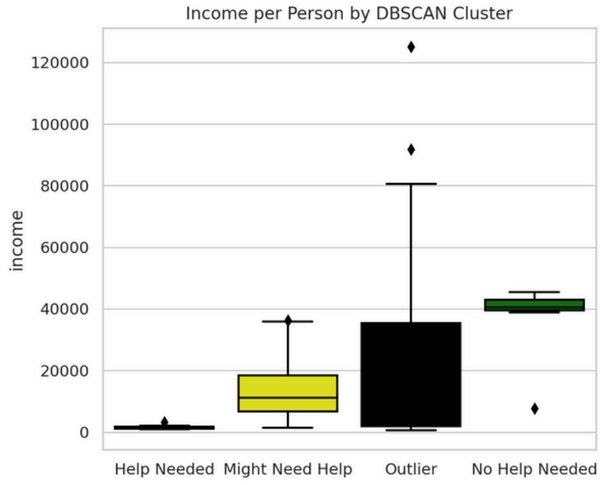
Metric	Feature-based (eps=0.08)	PCA-based (eps=0.20)	Better?
Clusters (excluding noise)	3	3	tied
Noise points	67	94	Feature-based (fewer)
Silhouette Score	**0.046**	−0.046	Feature-based
Calinski–Harabasz Score	**15.564**	8.390	Feature-based
Davies–Bouldin Index	**1.764**	2.733	Feature-based (lower)







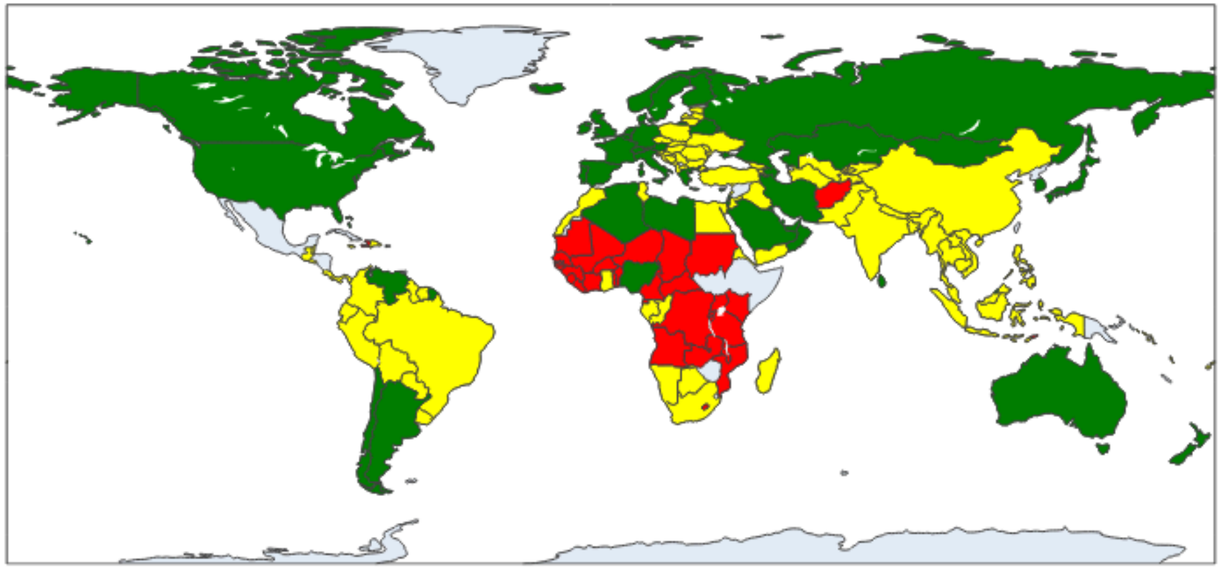


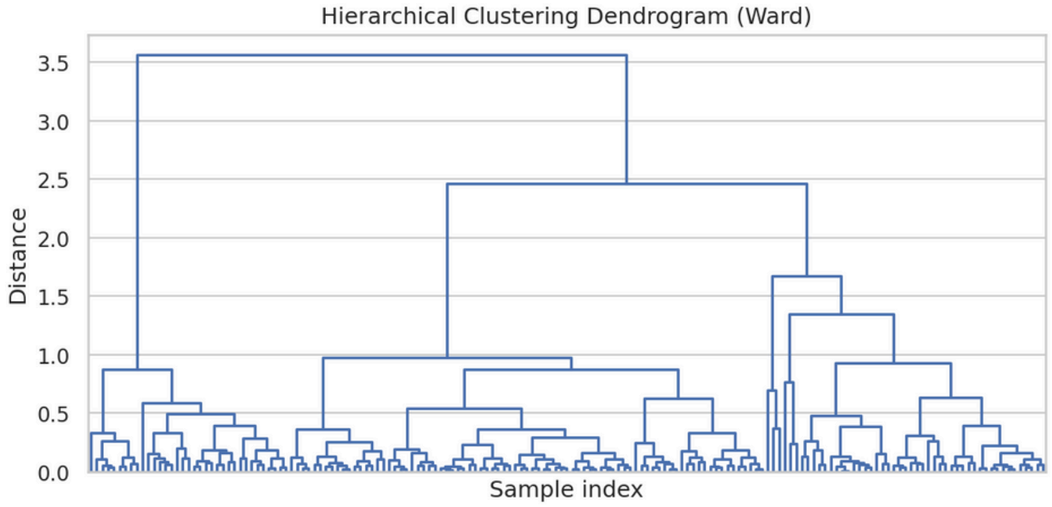


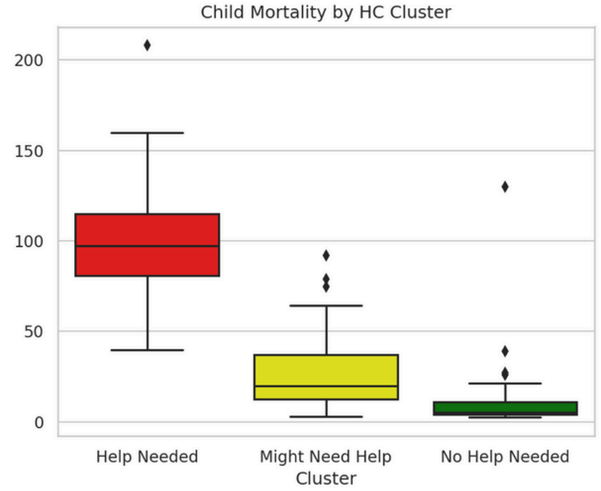
HIERARCHICAL CLUSTERING RESULTS

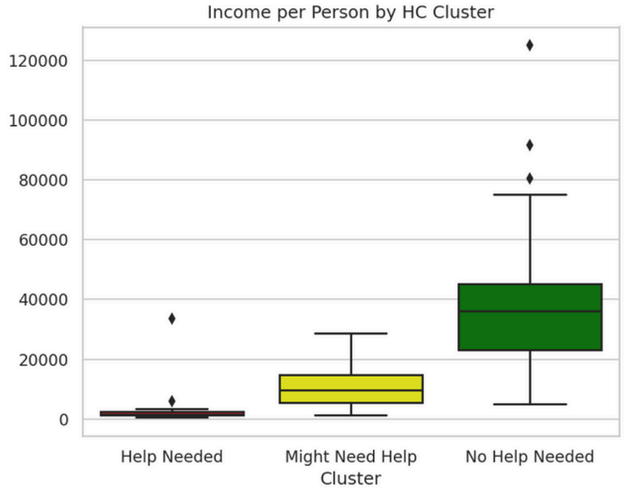
Quantitative Evaluation

Metric	Feature-based	PCA-based	Better?
Silhouette Score	0.378	**0.405**	PCA-based
Calinski–Harabasz Score	110.739	**186.003**	PCA-based
Davies–Bouldin Index	0.975	**0.810**	PCA-based (lower)
Cluster sizes	{0:49,1:35,2:83}	{0:94,1:24,2:49}	n/a









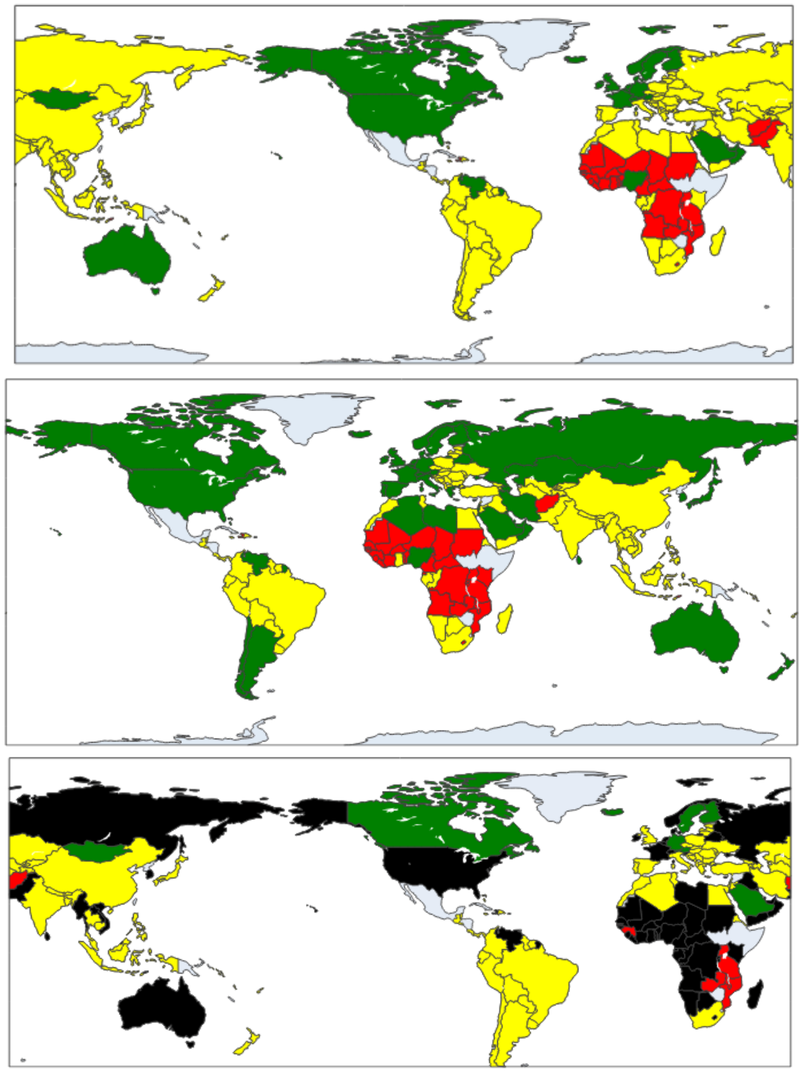
DISCUSSIONS & FINAL RECOMMENDATION

Summary of Findings

- **K-Means (Feature-based):**
  - Best silhouette score (0.452)
  - Clear, interpretable clusters aligned with real-world needs
- **Hierarchical Clustering (PCA-based):**
  - Strong metrics (CH score: 186.0, DB index: 0.810)
  - Useful when hierarchical structure is needed
- **DBSCAN:**
  - Lower clustering quality but valuable for outlier detection

Final Recommendation

- Use K-Means (Feature-based) as primary model
- Use Hierarchical Clustering for stakeholder presentation
- Use DBSCAN to detect extreme or unique country profiles



FUTURE WORK & OPPORTUNITIES

Data Improvements

- Add more features (e.g. education, digital access, governance)
- Use time-series for trend-based clustering

Feature & Model Enhancements

- Apply nonlinear transformations (log, Box-Cox)
- Try alternative scaling methods (e.g. RobustScaler)
- Test advanced models: GMM, HDBSCAN, Spectral

Interpretability & Validation

- Build interactive dashboards (e.g. Streamlit)
- Compare with UN HDI, World Bank groups
- Include expert review for domain validation


THANK YOU