Unsupervised Learning on Country Data

Strategic Aid Allocation for HELP International

Comparison of Machine Learning clustering algorithms on Country Dataset

Author: Matteo Vezzelli, PhD
https://www.linkedin.com/in/matteovezzelli/

International Aid Allocation

Unsupervised Learning on Country Data - Strategic Aid Allocation for HELP International

- Analysis of 167 countries
- $10 million humanitarian fund allocation
- Machine Learning approach: K-Means Clustering

Author: Matteo Vezzelli, PhD
https://www.linkedin.com/in/matteovezzelli/

Python Notebook available here: https://github.com/mtvz42/International-Aid-Allocation/tree/main

Problem Statement

Objective
HELP International needs to allocate $10 million strategically to countries in greatest need.

Challenge
How to identify priority countries objectively?

Approach
Unsupervised learning to categorize countries based on socio-economic and health indicators.

Dataset Overview

Source
https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/code

Scope
167 countries, 10 features, 9 development indicators

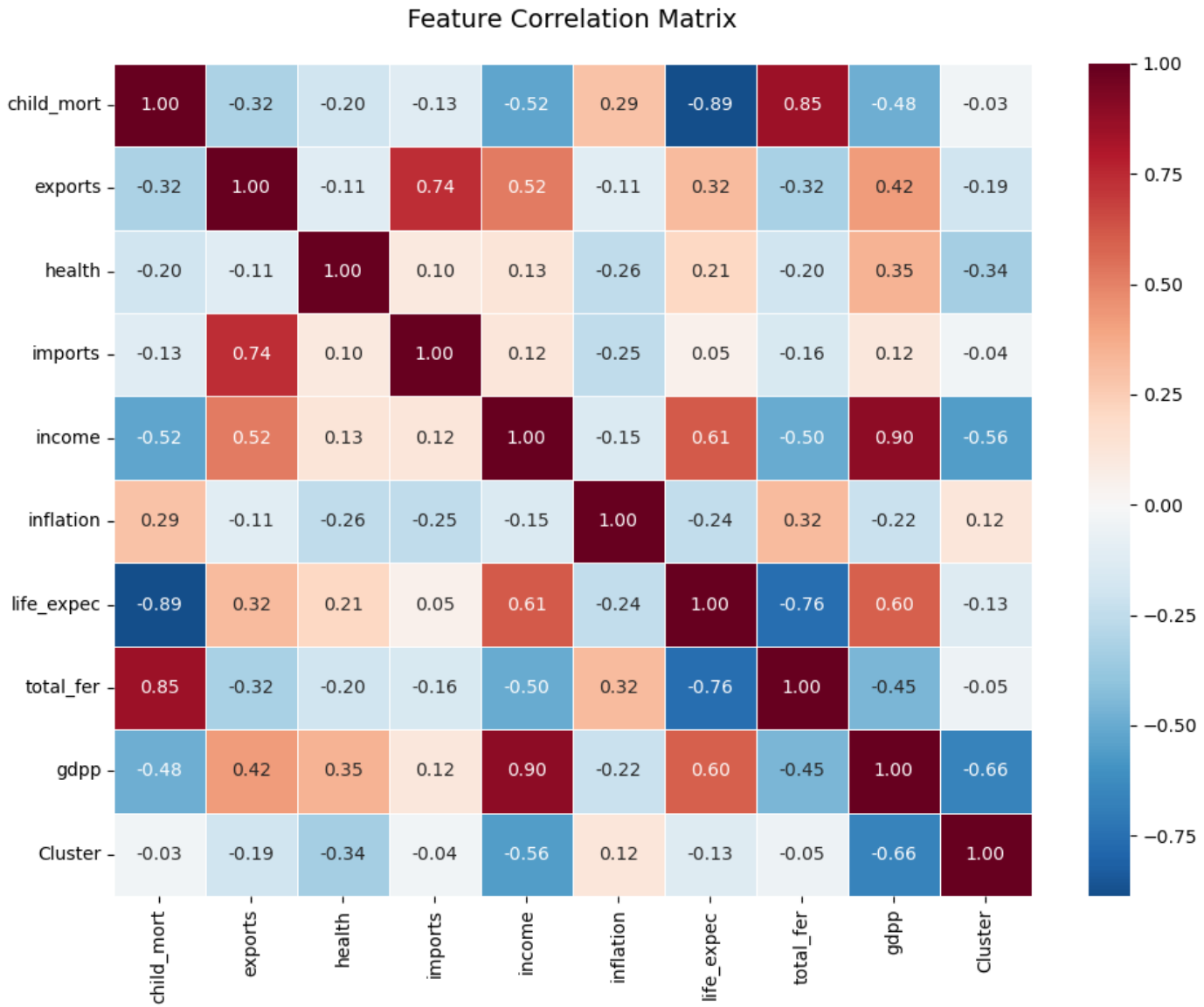
Features
- **country**: 167 countries
- **child_mort**: Child mortality rate (deaths per 1000 live births under age 5)
- **exports**: Exports as % of GDP per capita
- **health**: Health spending as % of GDP per capita
- **imports**: Imports as % of GDP per capita
- **income**: Net income per person (USD)
- **inflation**: Annual GDP growth rate (%)
- **life_expec**: Life expectancy at birth (years)
- **total_fer**: Fertility rate (children per woman)
- **gdpp**: GDP per capita (USD)

Key Features
- Health: child_mort, life_expec, health spending
- Economic: income, gdpp, exports, imports
- Demographic: total_fer, inflation

Dataset Characteristics:
- No missing values (167/167 complete)
- All numerical features (except country name)
- Appropriate data types (float64)
- No duplicates

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Correlation Analysis



Strong **Negative** Correlations:

- child_mort ↔ life_expec (r = -0.89)
- total_fer ↔ life_expec (r = -0.76)

Strong **Positive** Correlations:

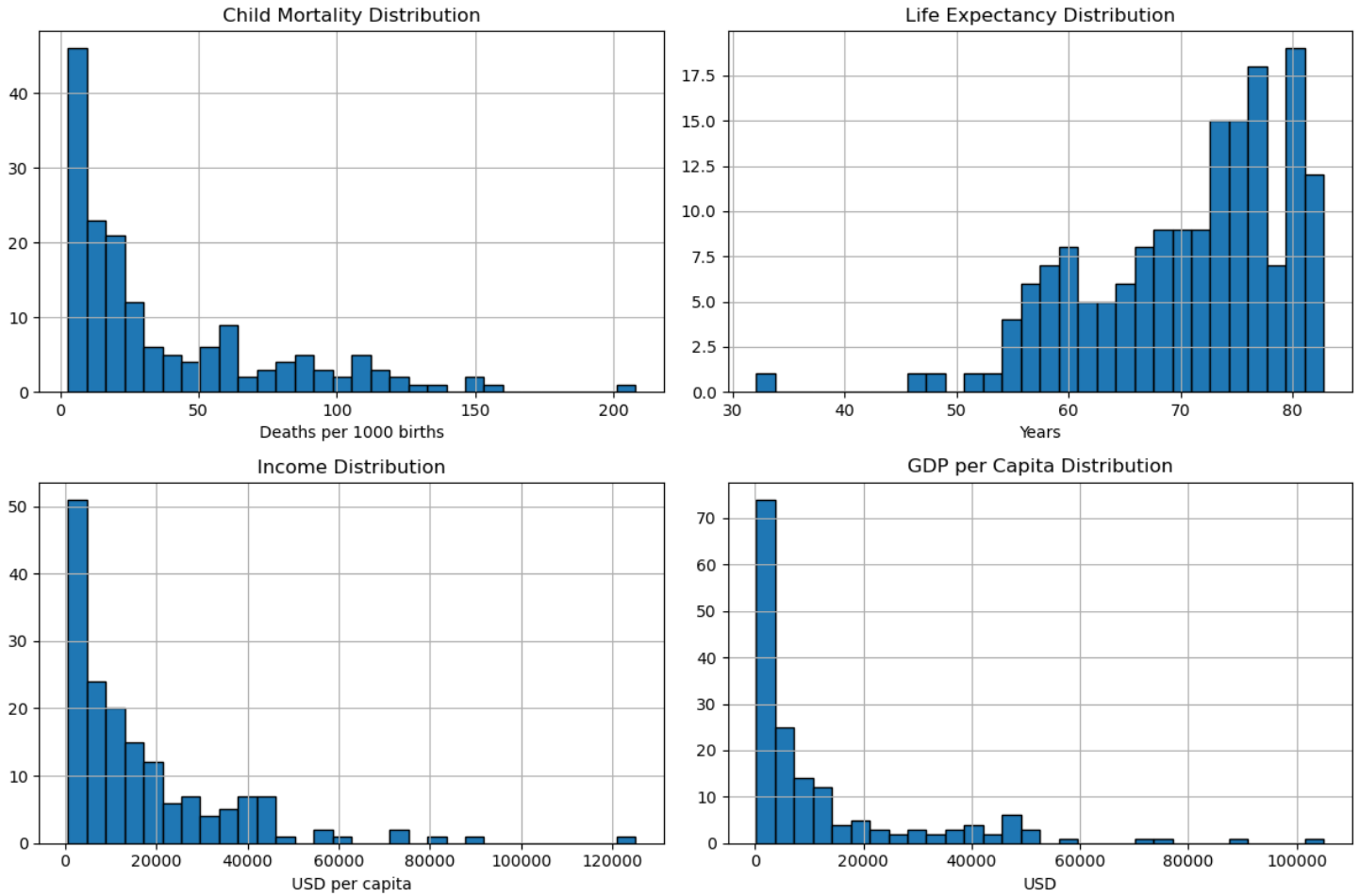
- income ↔ gdpp (r = 0.90)
- child_mort ↔ total_fer (r = 0.85)

These correlations suggest that child mortality, income, GDP, and life expectancy form the primary indicators of development status.

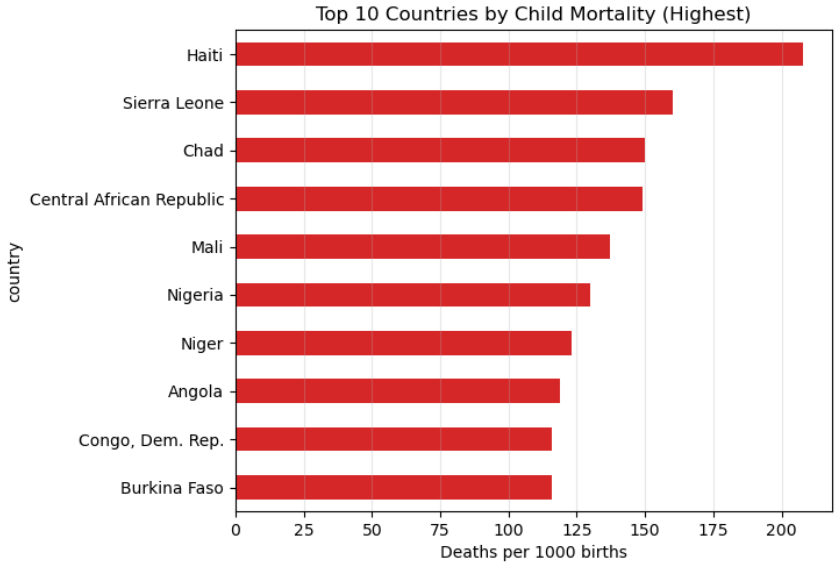
Feature Distributions

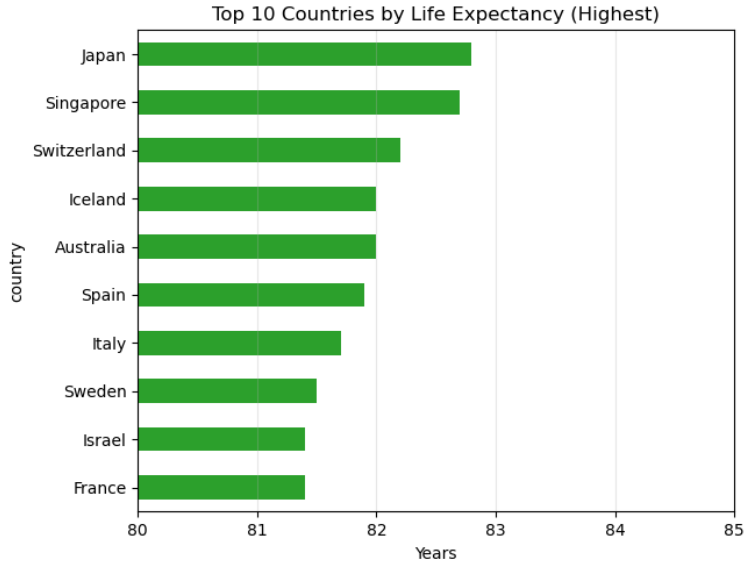
Key Observations:

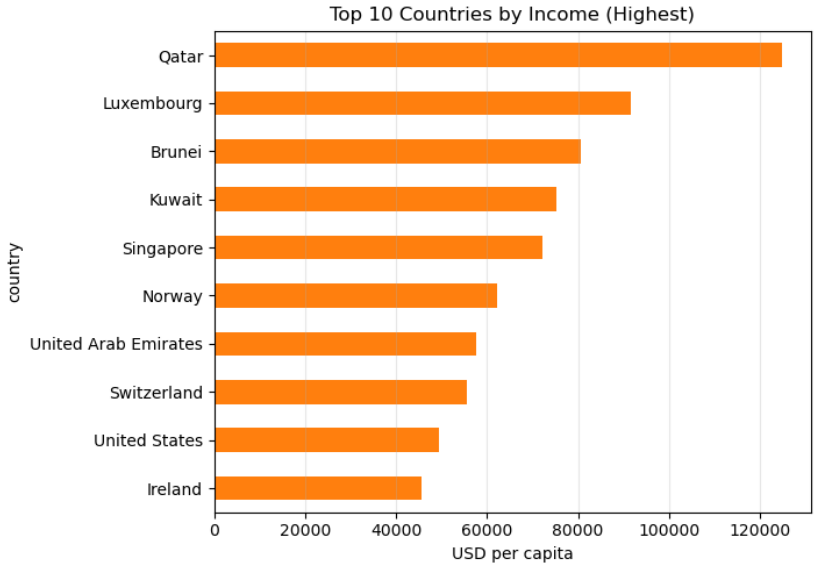
- Right-skewed distributions for income and GDP
- Wealth concentration in fewer nations
- High variability across all indicators
- Clear need for clustering

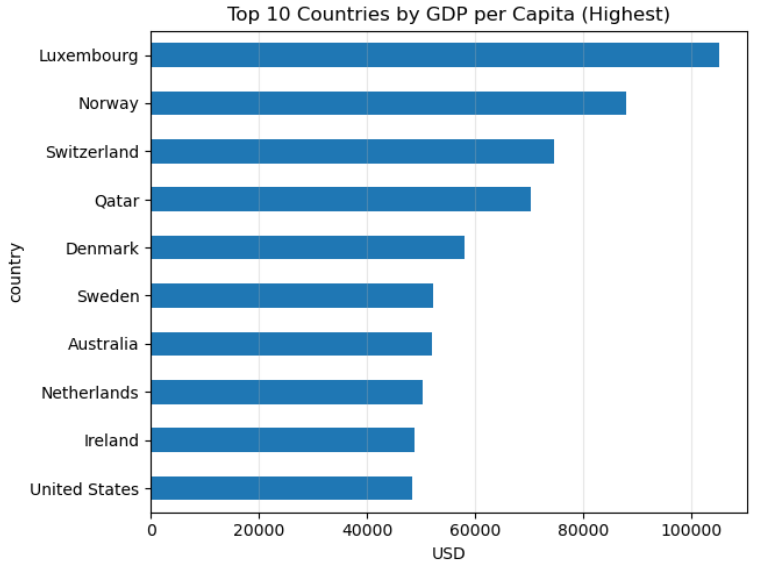


Extreme Values









Child Mortality (highest)
Average: >100 deaths per 1000 births

Life Expectancy (highest)
Range: 81-83 years

Income and GDP Disparities (highest)
Exceed $40,000 per capita

Data Preprocessing

Standardization required. Why?

- Features have vastly different scales (GDP: thousands, Inflation: percentages)
- Distance-based algorithms need equal feature contribution

Method

StandardScaler (mean=0, std=1)

Before scaling

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

After scaling

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanistan	1.291532	-1.138280	0.279088	-0.082455	-0.808245	0.157336	-1.619092	1.902882	-0.679180
Albania	-0.538949	-0.479658	-0.097016	0.070837	-0.375369	-0.312347	0.647866	-0.859973	-0.485623
Algeria	-0.272833	-0.099122	-0.966073	-0.641762	-0.220844	0.789274	0.670423	-0.038404	-0.465376
Angola	2.007808	0.775381	-1.448071	-0.165315	-0.585043	1.387054	-1.179234	2.128151	-0.516268
Antigua and Barbuda	-0.695634	0.160668	-0.286894	0.497568	0.101732	-0.601749	0.704258	-0.541946	-0.041817

Methodology Overview

Clustering Approach:

1. **K-Means** (primary)
- Partitions data into spherical clusters
- Minimizes within-cluster variance
- Best for evenly distributed data

2. **DBSCAN** (comparison)
- Density-based clustering
- Identifies outliers automatically
- Tests for non-spherical patterns

3. **Hierarchical** (comparison)
- Shows nested relationships
- No random initialization
- Validates K-Means results

Why Multiple Algorithms?
- Cross-validation of results
- Each algorithm has different assumptions
- Ensures clustering is not method-specific artifact

Evaluation: Silhouette Score
- Measures cluster quality (cohesion + separation)
- Range: -1 to +1 (higher = better)
- Objective metric for comparing algorithms

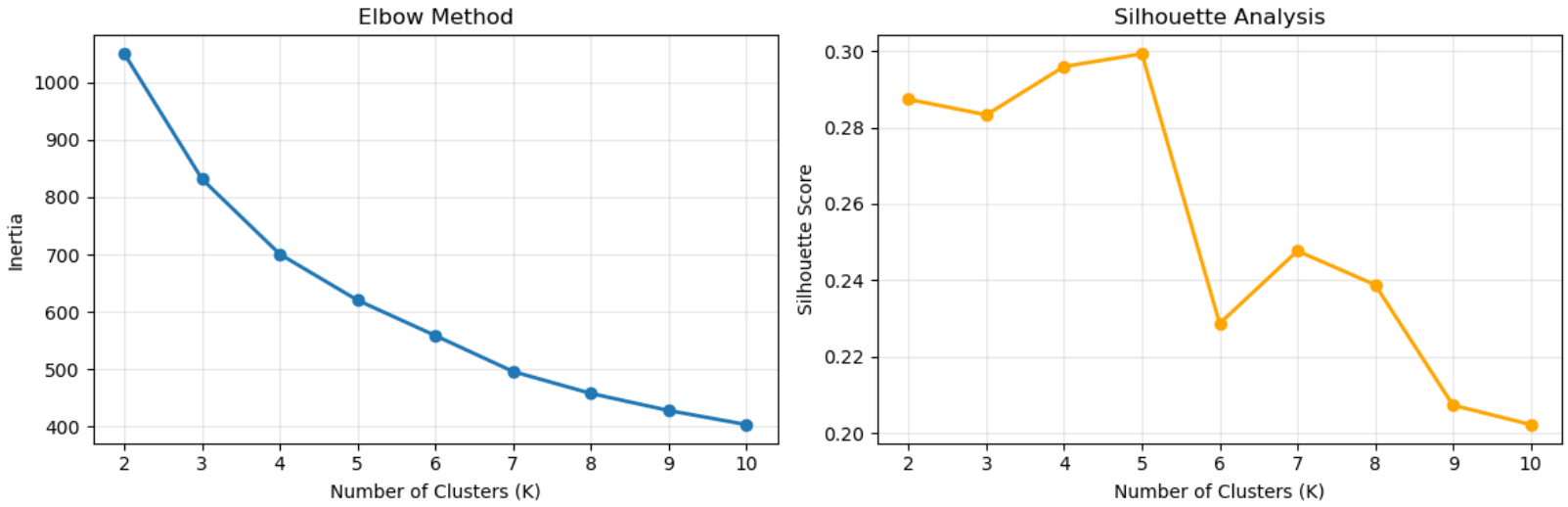
K-Means Algorithm

Elbow Method: Identify point of diminishing returns

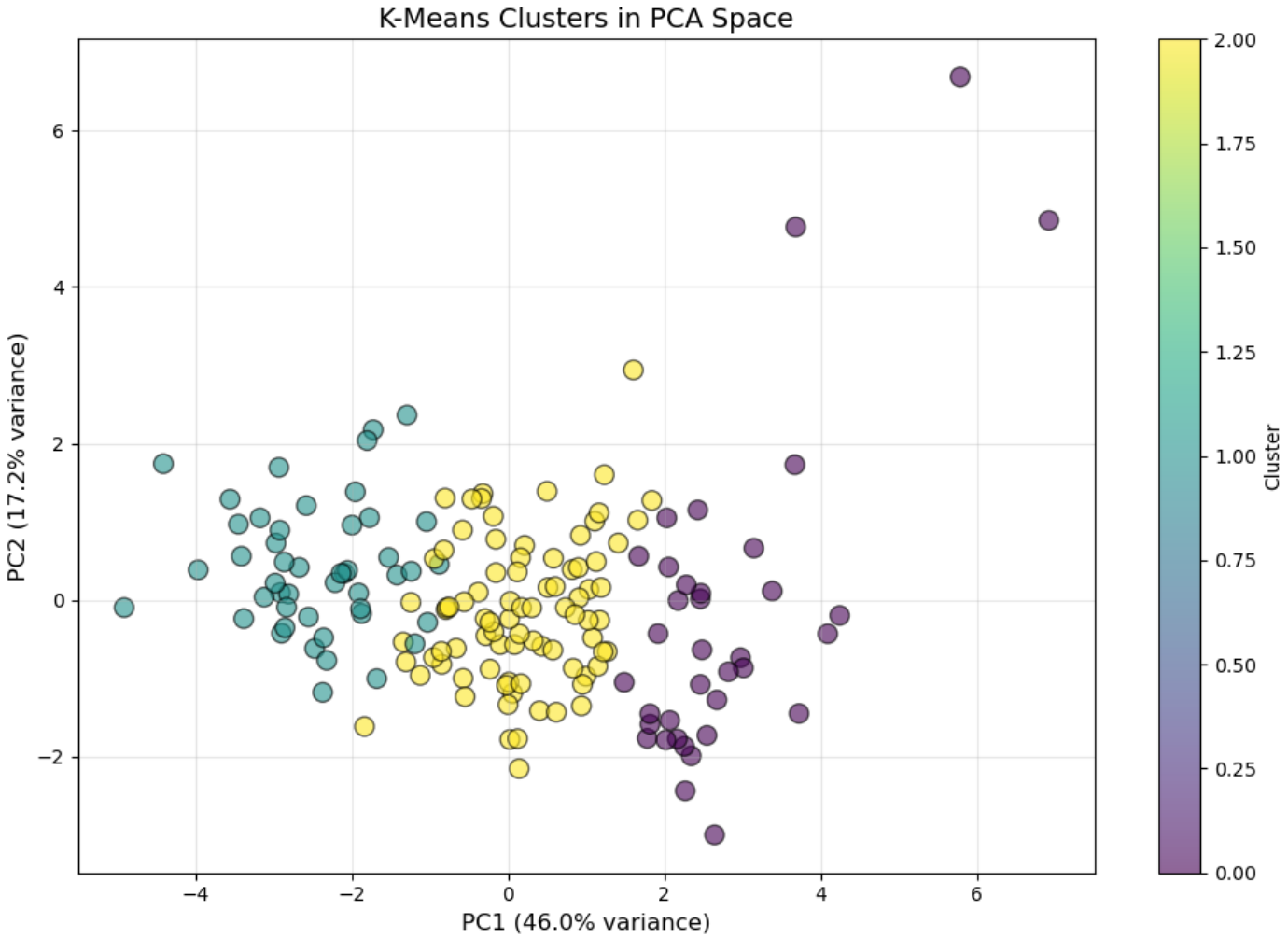
Silhouette Analysis: Measure cluster separation quality
Scores obtained: K=5 (0.299), K=4 (0.296), K=3 (0.283)

Choosen value: K=3 for the following reasons:

- The difference between K=5 and K=3 is only 0.016 (5.3% improvement)
- Good cluster separation even with K=3
- K=3 provides clear, actionable categories: Developed, Developing, Underdeveloped
- Balance between granularity and simplicity



PCA Visualization



Purpose:

- Visual validation of cluster separation
- Confirm distinct country groupings

Dimensionality Reduction: 9D → 2D

Variance Explained: ~63% (PC1 + PC2)

PC1: Development gradient
(underdeveloped → developed)

PC2: Secondary variation
(trade balance, resource distribution)

Observation:
Clear separation between clusters validates K=3 choice

Cluster Statistics

Three distinct groups identified:

Cluster 0 - Developed Nations: Low child mortality (<10 per 1000), high life expectancy (>75 years), substantial income (>$40,000), and high GDP. These nations are self-sufficient.

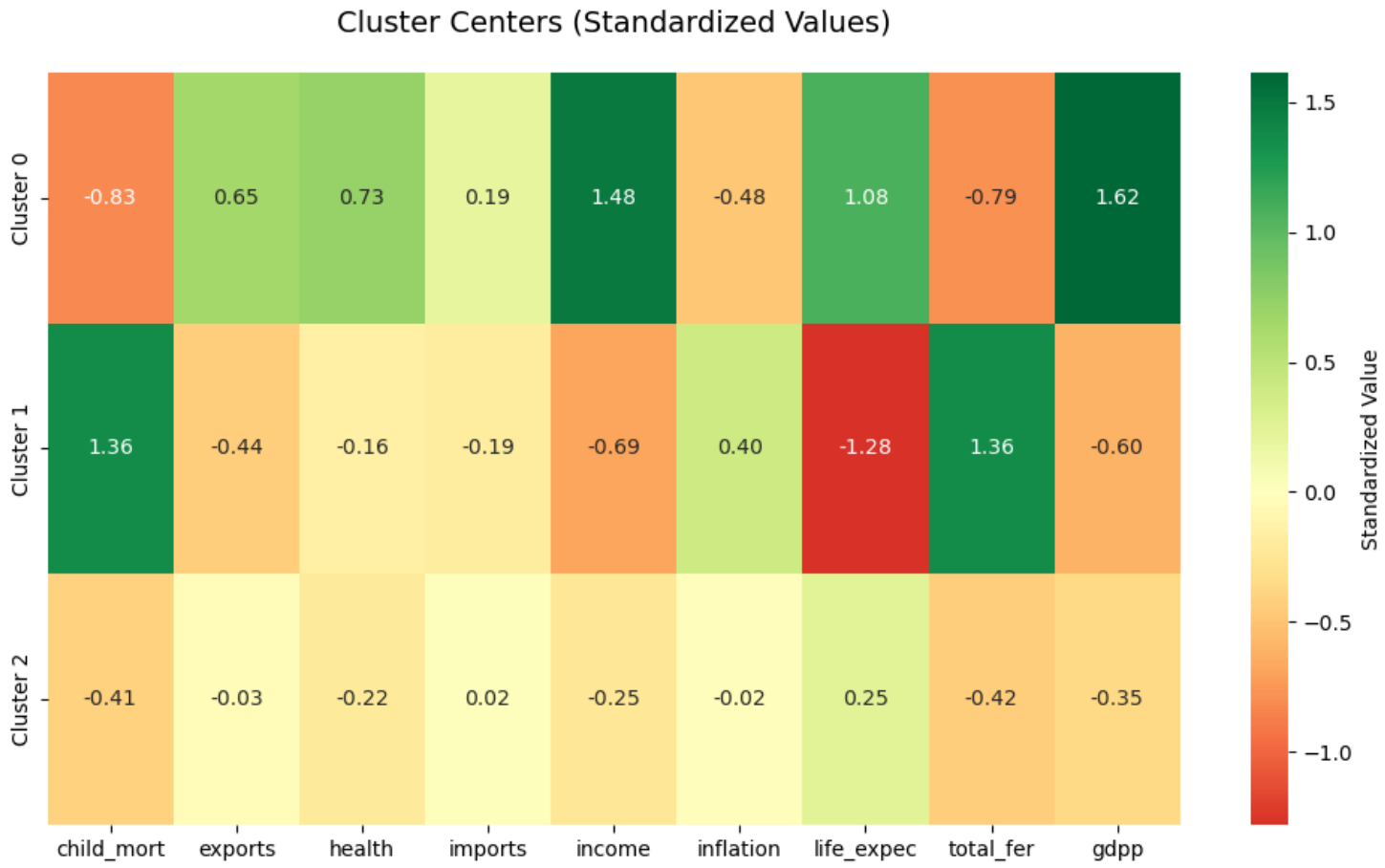
Cluster 1 - Underdeveloped Nations: Characterized by high child mortality (>80 per 1000), low life expectancy (<60 years), minimal income (<$5,000), and low GDP. These countries require immediate humanitarian intervention.

Cluster 2 - Developing Nations: Moderate indicators with child mortality between 20-80 per 1000, life expectancy 60-75 years, and GDP $5,000-$20,000. These countries are transitioning but need targeted support.

```
Cluster Profiles (Original Scale):
============================================================
         child_mort  life_expec    income      gdpp  health  total_fer  \
Cluster
0              5.00       80.13  45672.22  42494.44    8.81       1.75
1             92.96       59.19   3942.40   1922.38    6.39       5.01
2             21.93       72.81  12305.60   6486.45    6.20       2.31

         n_countries
Cluster
0                 36
1                 47
2                 84
============================================================
```

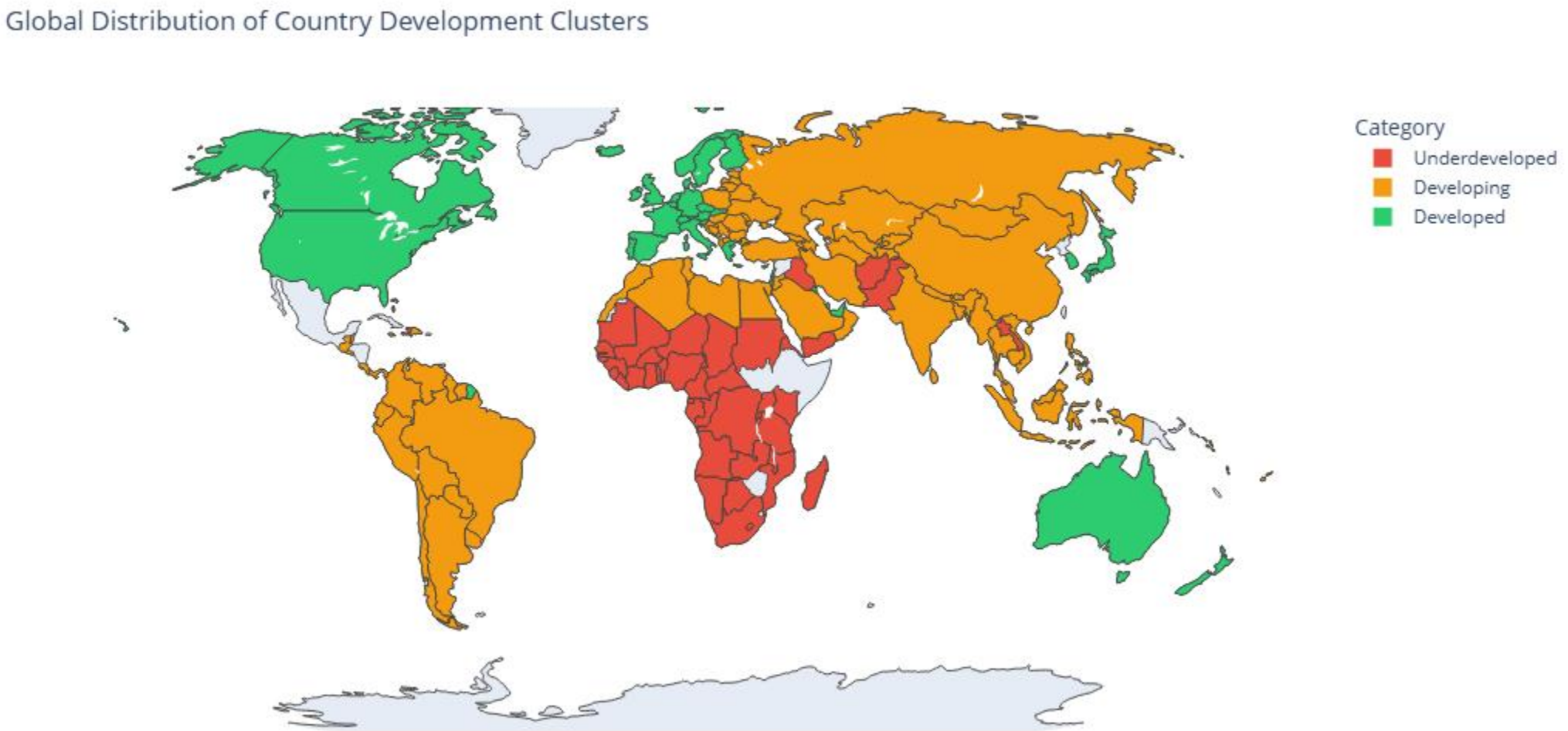
Cluster Centers Heatmap



Standardized Centroids of each cluster:

- **Green**: Above-average values
- **Red**: Below-average values
- **White**: Near global mean

Geographic Distribution



Priority Countries

```
High-Need Cluster: 1
Total countries requiring aid: 47

Top 15 Priority Countries:
================================================================================
 1. Haiti                      | Child Mort:  208.0 | Life Exp:  32.1 | GDP: $     662
 2. Sierra Leone               | Child Mort:  160.0 | Life Exp:  55.0 | GDP: $     399
 3. Chad                       | Child Mort:  150.0 | Life Exp:  56.5 | GDP: $     897
 4. Central African Republic   | Child Mort:  149.0 | Life Exp:  47.5 | GDP: $     446
 5. Mali                       | Child Mort:  137.0 | Life Exp:  59.5 | GDP: $     708
 6. Nigeria                    | Child Mort:  130.0 | Life Exp:  60.5 | GDP: $   2,330
 7. Niger                      | Child Mort:  123.0 | Life Exp:  58.8 | GDP: $     348
 8. Angola                     | Child Mort:  119.0 | Life Exp:  60.1 | GDP: $   3,530
 9. Burkina Faso               | Child Mort:  116.0 | Life Exp:  57.9 | GDP: $     575
10. Congo, Dem. Rep.           | Child Mort:  116.0 | Life Exp:  57.5 | GDP: $     334
11. Guinea-Bissau              | Child Mort:  114.0 | Life Exp:  55.6 | GDP: $     547
12. Cote d'Ivoire              | Child Mort:  111.0 | Life Exp:  56.3 | GDP: $   1,220
13. Benin                      | Child Mort:  111.0 | Life Exp:  61.8 | GDP: $     758
14. Equatorial Guinea          | Child Mort:  111.0 | Life Exp:  60.9 | GDP: $  17,100
15. Guinea                     | Child Mort:  109.0 | Life Exp:  58.0 | GDP: $     648
```

Ranked by child mortality rate (highest to lowest)

Alternative Algorithm - DBSCAN

Density-Based Clustering

How it works: Groups high-density regions, marks outliers

Result: Underperformed (Silhouette: varies)

Why?
Country development follows a continuous gradient without clear density gaps. DBSCAN excels at finding isolated clusters (e.g., customer segments), but our data lacks these distinct separations. Result: either one massive cluster or excessive noise points.

Business impact: We need to classify ALL countries for aid allocation. DBSCAN's outlier detection leaves some countries unclassified, making it unsuitable for our purpose.

```python
# DBSCAN with parameter search
best_score = -1
best_params = None

for eps in np.arange(2.0, 4.5, 0.5):
    for min_samples in [4, 5, 6]:
        dbscan = DBSCAN(eps=eps, min_samples=min_samples)
        labels = dbscan.fit_predict(data_scaled)

        n_clusters = len(set(labels)) - (1 if -1 in labels else 0)
        if n_clusters >= 2:
            mask = labels != -1
            if mask.sum() > 0:
                score = silhouette_score(data_scaled[mask], labels[mask])
                if score > best_score:
                    best_score = score
                    best_params = (eps, min_samples, n_clusters)

if best_params:
    print(f"Best DBSCAN: eps={best_params[0]:.1f}, min_samples={best_params[1]}")
    print(f"Clusters found: {best_params[2]}, Silhouette: {best_score:.3f}")
    dbscan_score = best_score
else:
    print("DBSCAN: No valid clustering found")
    dbscan_score = -1.0
```

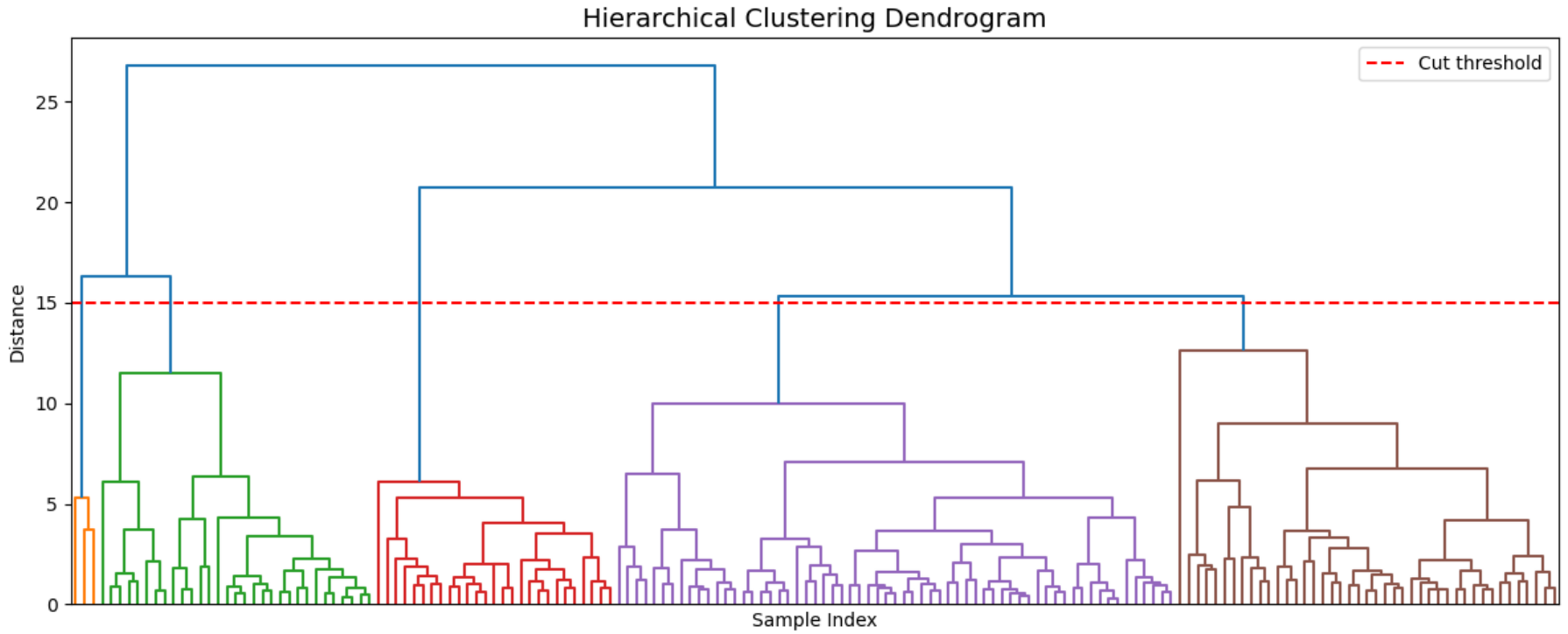
✓ 0.1s

```
DBSCAN: No valid clustering found
```

Alternative Algorithm - Hierarchical

Agglomerative Clustering

How it works: builds tree by merging similar clusters
Result: similar to K-Means (Silhouette: 0.246)
Advantage: shows nested relationships
Limitation: less practical for flat categorization



Model Comparison

Model Selection: K-Means is selected as the final model based on:

1. **Highest silhouette score** indicating well-separated clusters
2. **Complete country coverage** enabling comprehensive aid allocation
3. **Interpretable centroids** providing clear cluster definitions
4. **Computational efficiency** and reproducibility

DBSCAN's lower performance suggests country development exists on a continuous spectrum rather than as distinct density-based groups.
Hierarchical clustering produces similar results but lacks explicit cluster centers.

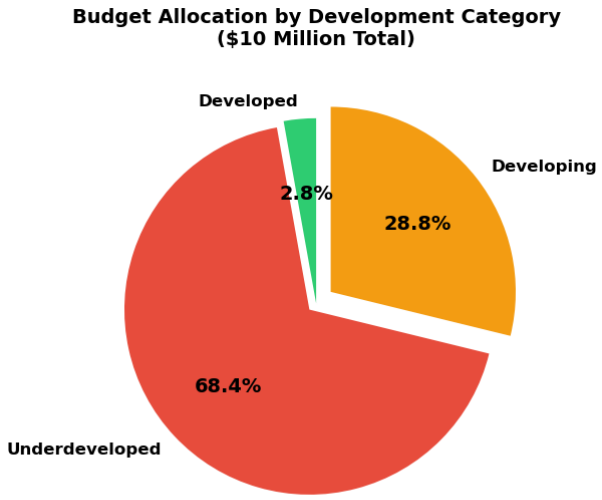
```
Algorithm Performance Comparison:
============================================================
   Algorithm  Silhouette Score  N Clusters Complete Coverage
     K-Means          0.283296           3               Yes
Hierarchical          0.245630           3               Yes
      DBSCAN         -1.000000           0                No
============================================================
```

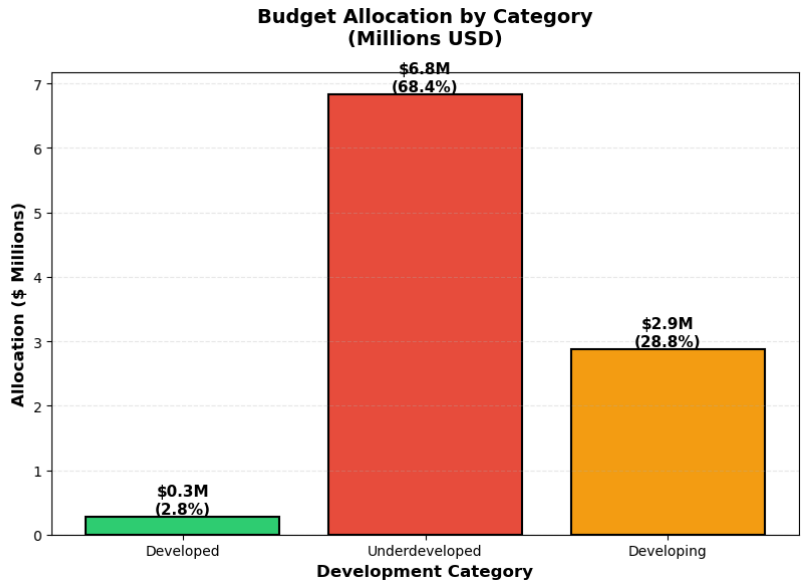
Strategic Fund Allocation

```
================================================================================
STRATEGIC FUND ALLOCATION - $10,000,000
================================================================================
 Cluster       Category  Countries  Avg Child Mortality  Need Score  Allocation ($)  Per Country ($)  Percentage (%)
       0      Developed         36                 5.00       180.0          281635             7823             2.8
       1 Underdeveloped         47                92.96      4369.0         6836107           145449            68.4
       2     Developing         84                21.93      1842.0         2882258            34313            28.8
================================================================================

Total Allocated: $10,000,000

Allocation Strategy:
- Proportional to (Child Mortality Rate × Number of Countries)
- Ensures both severity and scale are considered
- Developed nations (Cluster 0) receive minimal/no funding
```





Conclusions

Key Findings

- **Three distinct development clusters** identified among 167 countries
- **High-need cluster** contains countries with child mortality rates exceeding 80 per 1000 births
- **Strong correlation** between economic indicators (income, GDP) and health outcomes (child mortality, life expectancy)
- **K-Means clustering** outperforms alternative algorithms for this dataset

Limitations

- Analysis limited to socio-economic indicators; political stability and governance quality not considered
- Static dataset does not capture temporal changes
- Equal weighting of features may not reflect true importance for aid effectiveness
- Outliers within clusters may require individual assessment

Conclusion

This unsupervised learning analysis successfully identified three distinct country development categories, enabling data-driven allocation of humanitarian aid. K-Means clustering with K=3 provides the optimal balance of statistical performance and practical interpretability. The results align with established development classifications and offer actionable insights for HELP International's strategic decision-making.

Future work should incorporate additional variables (political stability, corruption indices), explore temporal dynamics, and validate recommendations through cost-benefit analysis.

Thank You!

Contact:
Matteo Vezzelli, PhD
https://www.linkedin.com/in/matteovezzelli/

Full Python notebook:
https://github.com/mtvz42/International-Aid-Allocation/tree/main