Revisiting the First–Third World Divide: A Data-Driven Classification of Global Development

Jack Bogner

2025-06-14

Abstract

During the Cold War, countries were classified as First World (NATO and its allies), Second World (the USSR and its allies), and Third World (non-aligned nations). In recent times, this terminology has been adapted to refer to an often generalized developmental divide. The terms "First World" and "Third World" are often used to distinguish between developed and developing countries. In reality, development is a complex and multifaceted concept that spans economic performance, governance quality, social welfare, infrastructure, and more. Modern indices such as the Human Development Index (HDI) aim to capture this complexity in a single metric, but many global disparities remain obscured by simplistic labels.

This project uses unsupervised machine learning techniques to group countries based on a set of economic and social indicators. Drawing from data provided by the World Bank and United Nations Development Program, it constructs a data-driven map of global development patterns. This project aims to evaluate the effectiveness of various machine learning techniques in constructing a generalized model of global developmental divisions among nations. By comparing clustering outcomes to conventional classifications, the project aims to provide insights into how developmental divides are shaped—and occasionally misrepresented—by popular narratives.

Introduction

Today, the terms "First World" and "Third World" are often used—though imprecisely—to distinguish between developed and developing countries. While these Cold War-era dichotomies are misnomers in the context of development, and such binary labels often oversimplify the nuanced realities of global development, which is shaped by diverse economic, social, and structural factors.

This project adopts a data-driven perspective to explore global development patterns through unsupervised machine learning. Specifically, it uses techniques—including K-means, Uniform Manifold Approximation and Projection (UMAP), Gaussian Mixture Models, and hierarchical agglomerative clustering—to group countries based on a suite of indicators related to economic activity, health, infrastructure, and demographics. These include GDP, population, exports, energy use, urbanization rate, health expenditure, internet accessibility, and more, sourced from the World Bank. Additionally, Human Development Index (HDI) data from the UNDP serves as a benchmark for evaluating developmental outcomes.

The analysis proceeds in two stages. First, clustering is used to generate a simplified developmental divide between "developed" and "developing" nations. This binary framing serves as a bridge between historical terminology and modern quantitative analysis. Second, the project moves beyond this two-cluster framework, applying each technique to produce a broader range of clusters. These outcomes are then compared and synthesized using an ensemble approach to identify areas of consensus and divergence among methods.

Through this process, the project aims not only to reconstruct a familiar global divide but also to challenge and refine it. By layering multiple analytical perspectives, it seeks to uncover regions that defy conventional classification—outliers, intermediates, and transitional economies that complicate the developed/developing dichotomy.

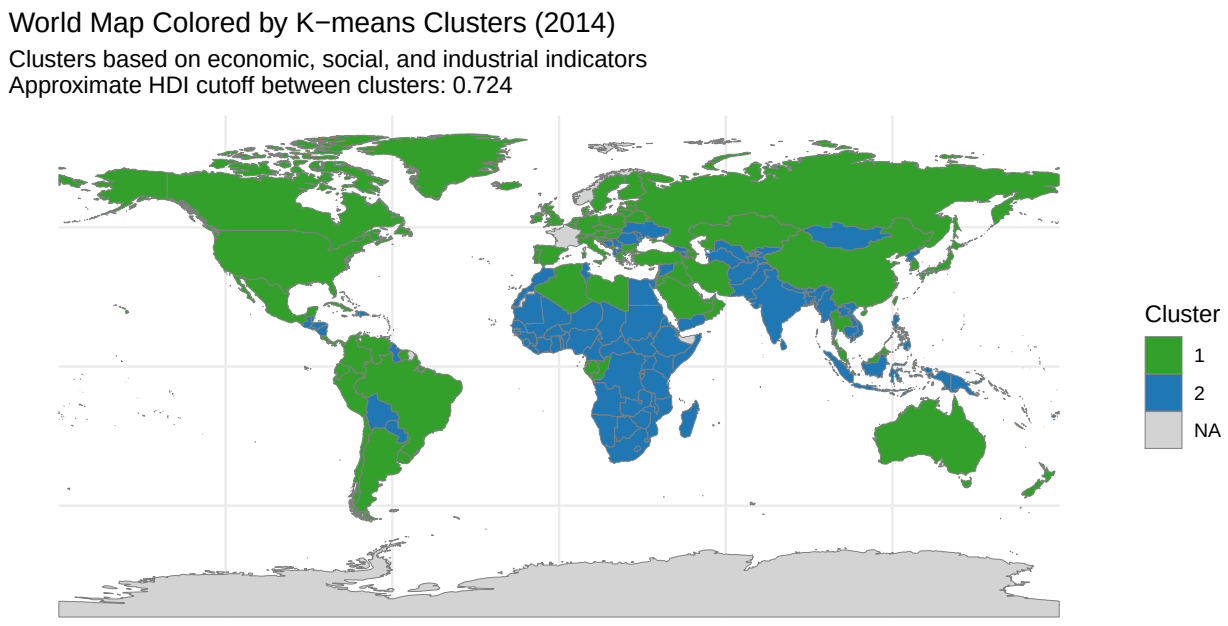
Data

Economic and social indicators will be sourced from World Bank data sets. Although the Human Development Index (HDI) serves as a social indicator, I use it primarily as a rough measure of a country's development level. HDI is preferred because it incorporates multiple factors and is less prone to distortion than single metrics like GDP per capita, which can be misleading for resource-rich countries with low living standards. HDI data is from the United Nations Development Program (UNDP).

Methods

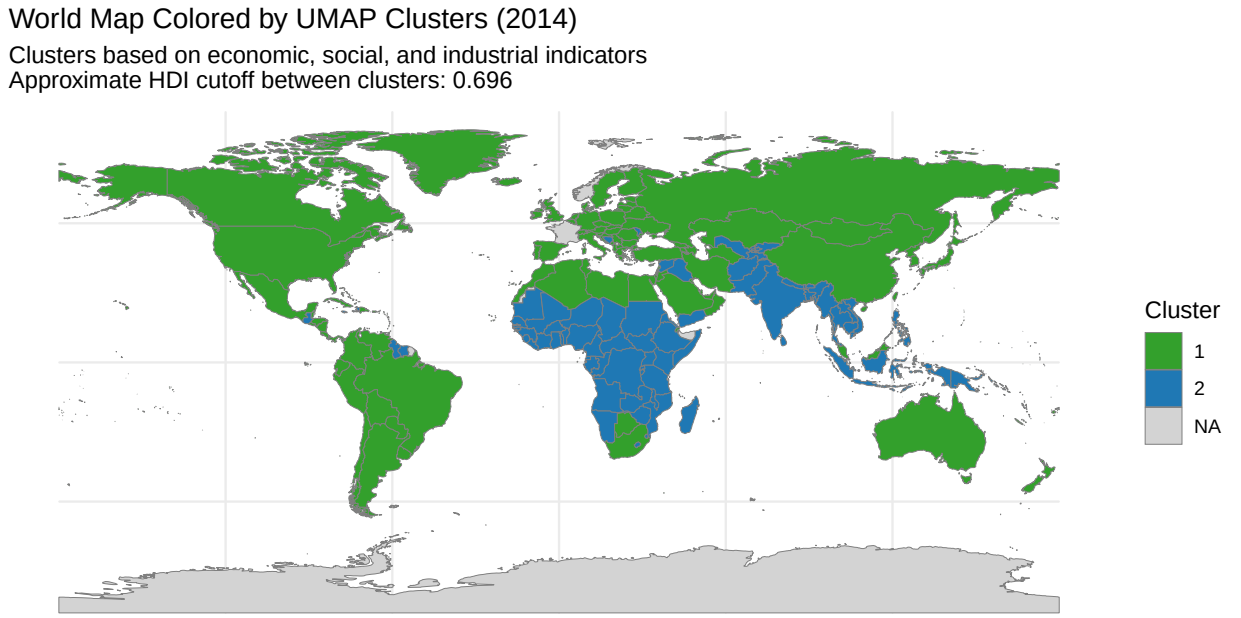
We will apply several clustering techniques but focus on those that allow us to specify the number of clusters, k. Initially, I set $k = 2$ to distinguish between "developed" (or "first world") and "developing" (or "third world") countries. The methods used include K-Means Clustering, Uniform Manifold Approximation and Projection (UMAP), Gaussian Mixture Models (GMM), and Hierarchical Clustering. Additionally, I will combine the results of these methods to assess whether an ensemble approach yields more accurate clustering. First, I created a map and a bar plot for each method to visualize regional cluster assignments and examine their relationship with HDI scores. Next, I evaluated which clustering methods best capture differences in HDI scores and increase the number of clusters (k) to explore intermediate and outlier regions. Exploratory analysis shows that 2014 has the most complete data. To impute missing values, I used K-Nearest Neighbors (KNN), drawing on data from 2012 to 2016 for each country or region.

Results

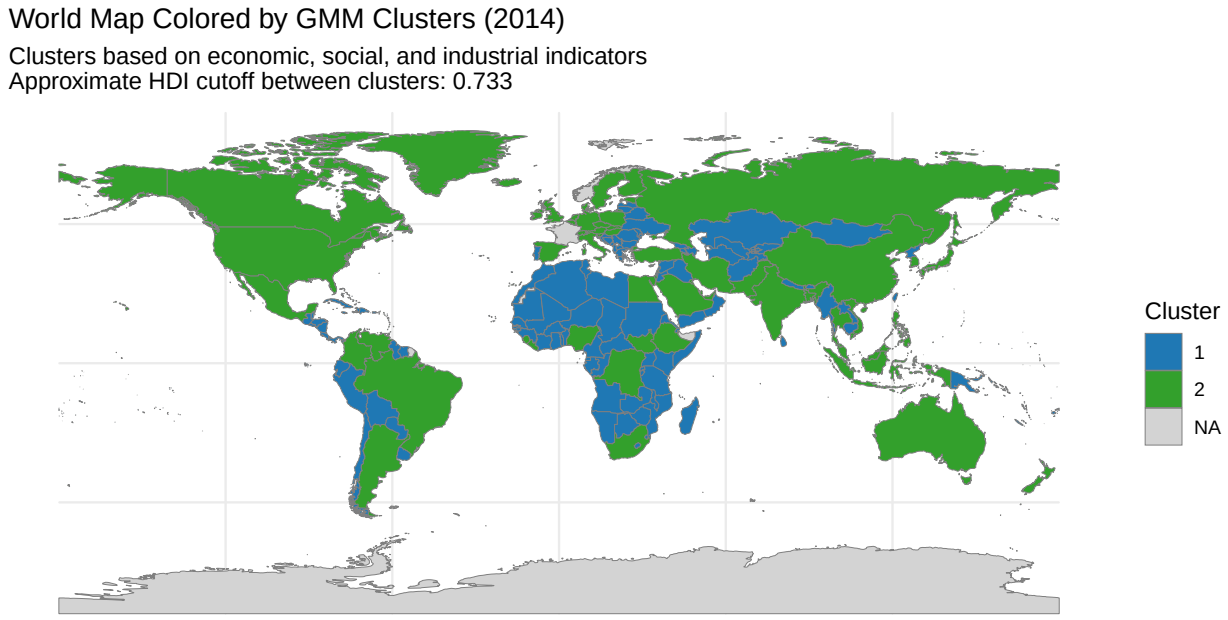


K-means is a clustering algorithm that partitions a dataset into k predefined groups by minimizing the Euclidean distance between data points and their respective cluster centroids. The algorithm iteratively updates cluster assignments and centroid positions until convergence. In our application, K-means produced

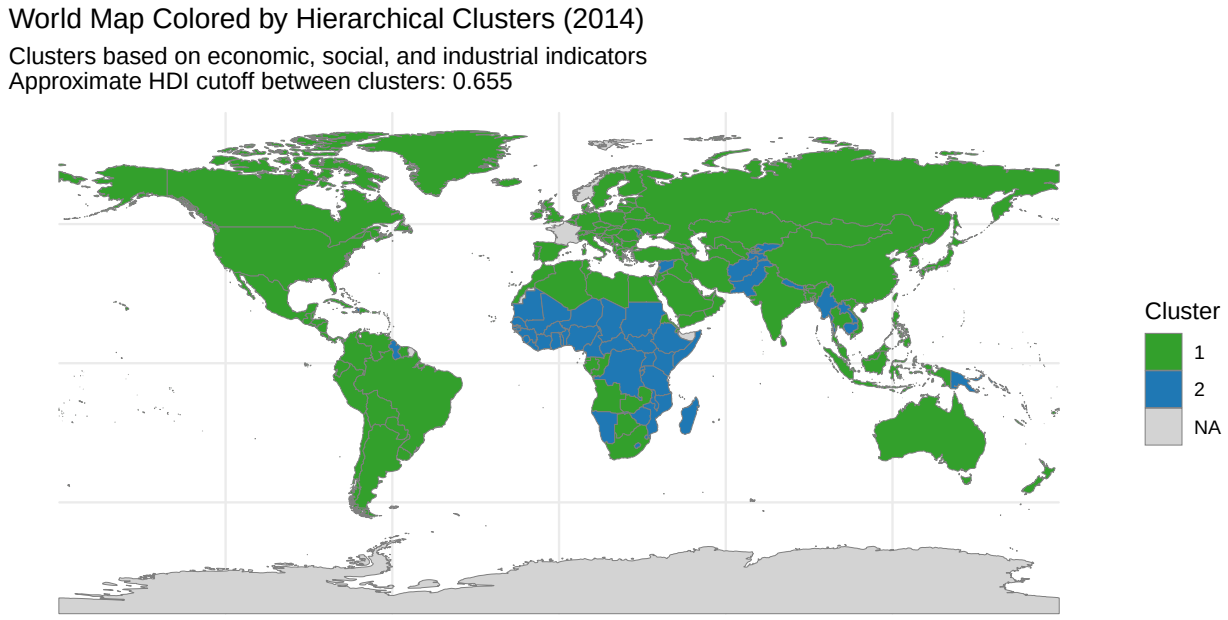
a relatively balanced split, assigning 111 countries to cluster 1 and 106 to cluster 2. However, K-means does not necessarily make conservative or intuitive assignments. For instance, most of Africa, along with parts of Central Asia and the Caucasus, fall into cluster 2. This includes countries like South Africa, India, Ukraine, and the Dominican Republic—nations often considered more developed than others in the same cluster, such as Syria, Chad, or the Democratic Republic of the Congo (DROTC). This illustrates a key limitation of K-means: it relies solely on numerical similarity in feature space, which may not always align with human or contextual interpretations of concepts like development.



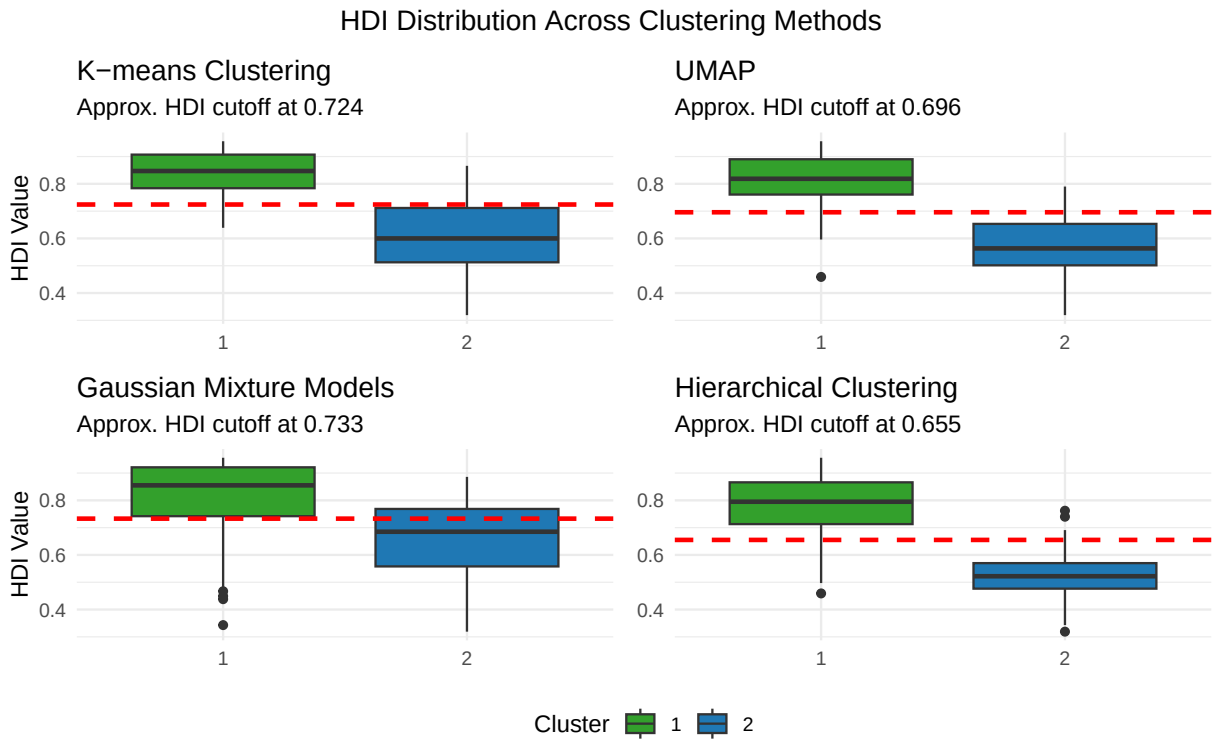
Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction technique which excels and preserving both local and global data structure. It constructs a graph based on the nearest neighbors in the original space, and then optimizes a low-dimensional layout which maintains this structure. In our application, UMAP was followed by K-means clustering, which groups the reduced data into two clusters. Cluster 1 is notably larger than it was using strictly K-means, with 136 countries belonging to cluster 1, and 81 belonging to cluster 2, and with this change UMAP produced some notable differences. Countries such as South Africa, Honduras, and North Korea have now been assigned to cluster 1, which is seemingly the more "developed" cluster, while Iraq, Thailand, and Gabon have shifted to cluster 2. These changes highlight the influence of dimensionality reduction on clustering outcomes, as well as the importance of interpreting clusters in the light of the methods used to generate them.



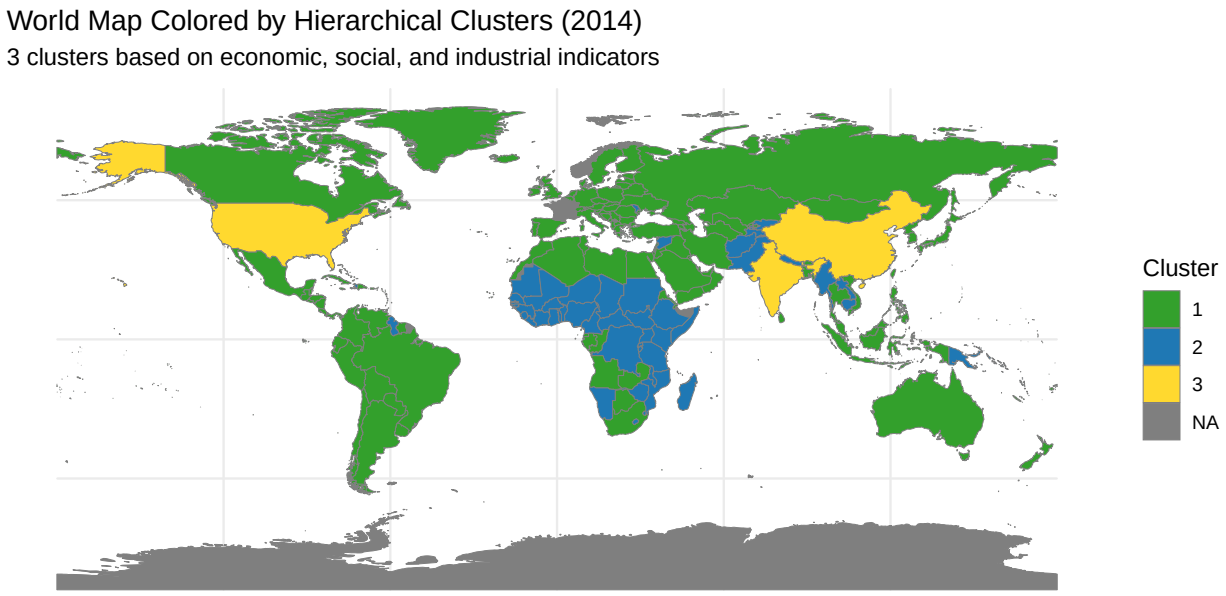
Gaussian Mixture Models (GMM) provide a probabilistic approach to clustering by assuming the data is generated from a mixture of multiple Gaussian distributions. Unlike K-means and UMAP, which assigns observations to clusters based on distance, GMM evaluates the probability of membership in each cluster, allowing for more flexible boundaries between groups. In our application, while cluster sizes are comparable to UMAP (132 in cluster 1, and 85 in cluster 2), actual assignments vary drastically. GMM has flipped the Caucasus back to cluster 2, along with large chunks of Latin and South America. Additionally, it has flipped India, Pakistan, and DROTC to cluster 1. This assignment change is reasonable for an economic powerhouse like India, though it is hard to justify the change in assignment for the DROTC, which is often rated among the lowest countries for human development. These changes reflect how GMM's probabilistic modeling can lead to a different interpretation of country groupings, particularly in cases where data does not conform to the hard, spherical data assumptions of K-means. Based solely on observation, it seems GMM is one of the weaker clustering methods in the context of our data.

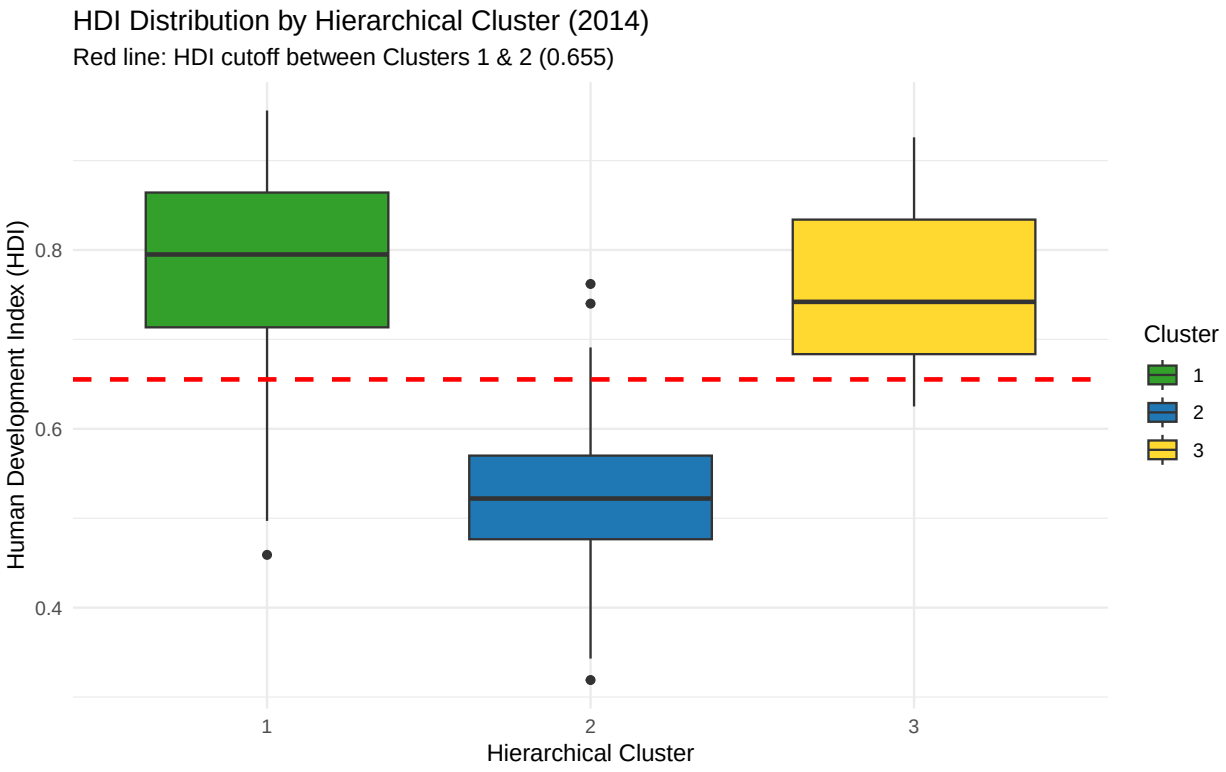


Hierarchical Clustering builds a multilevel hierarchy of clusters by either successively merging smaller clusters or splitting larger ones. Unlike K-means or GMM, it does not specify the number of clusters, and produces a dendrogram that helps visualize nested groupings and relative similarity between observations. I have cut this dendrogram at two groups, with 165 countries in cluster 1, and 52 in cluster 2. Notably, cluster 2 is centered in Africa, with notable mentions in Central and Southeast Asia. This clustering aligns well with global patterns of development, suggesting a division that meaningfully captures low and high levels of human development. Hierarchical clustering appears well-suited for our data, and it avoids some limitations of centroid-based methods such as K-means, which can overgeneralize clusters with broad or complex feature patterns. In this case, the hierarchical approach seems to reflect human development variation with more sensitivity.



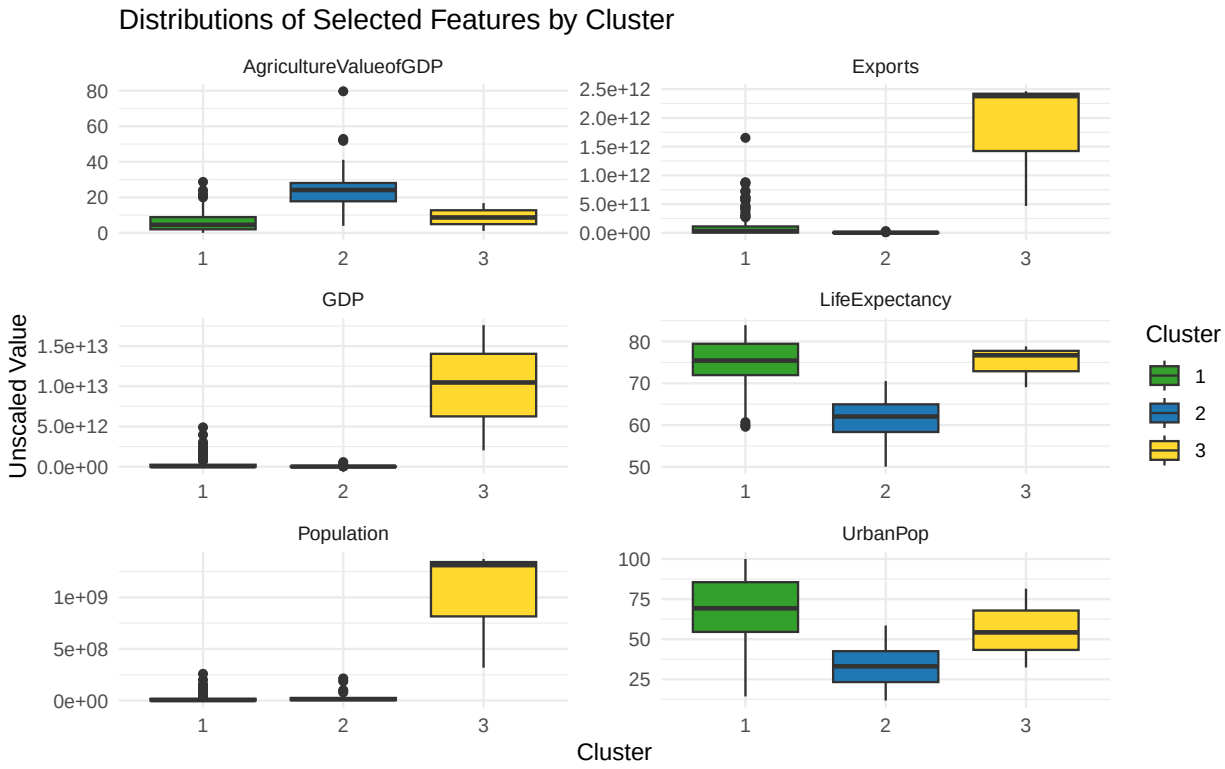
Comparing Human Development Index (HDI) values for each of our previous clustering methods, it is easier to visualize which methods best distinguish between low and high human development. The average of HDI means within groups is visualized in red to help visualize the degree of separation between clusters. From this perspective it is clear that UMAP and Hierarchical Clustering are better at forming groups which reflect human development, and they have notably smaller HDI cutoff values. Interestingly, the methods which have a greater degree of separation also estimate smaller HDI cutoffs. The cutoffs for UMAP and Hierarchical Clustering are both less than the 2014 HDI mean of 0.714. While this is not enough evidence to draw any conclusions about an ideal HDI cutoff value, it does suggest that methods producing more polarized groupings may be better aligned with real-world development disparities. I can say with confidence that Hierarchical Clustering is the best suited method for our data, and has the highest number of "reasonable" assignments. For investigative purposes, I increased the number of clusters k to three.

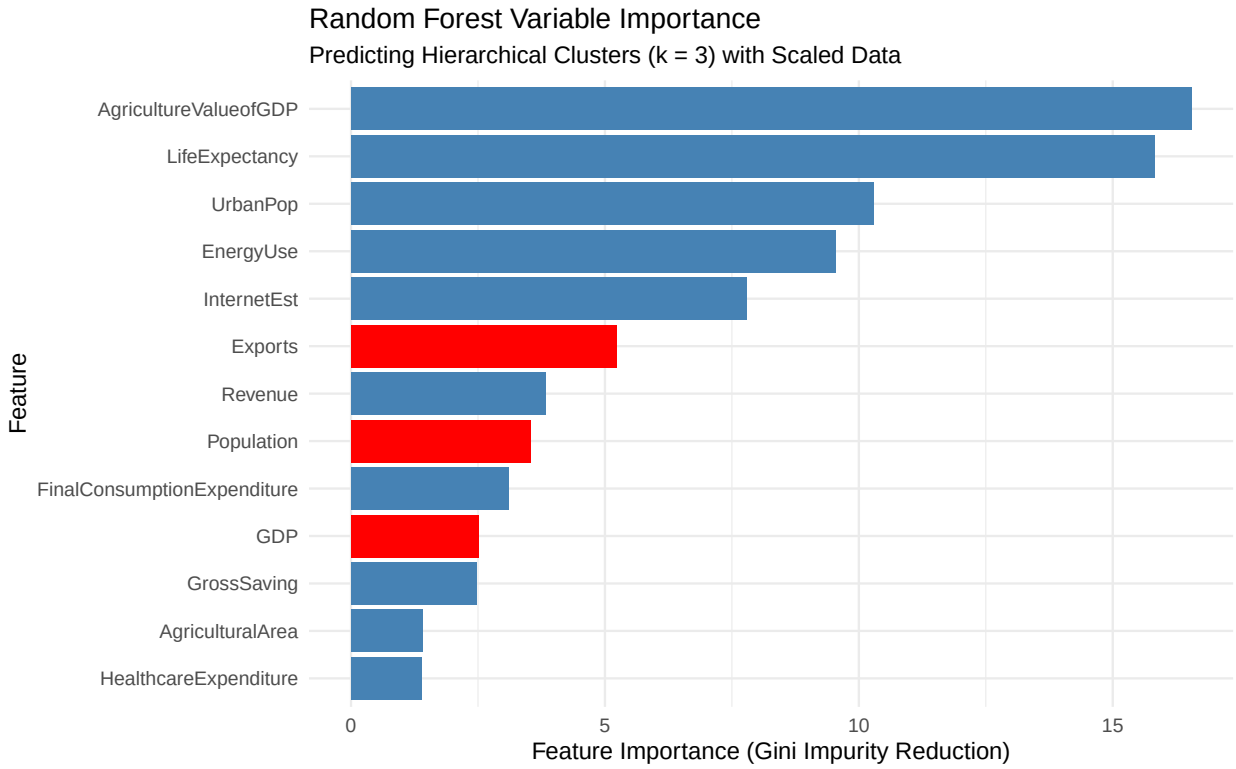


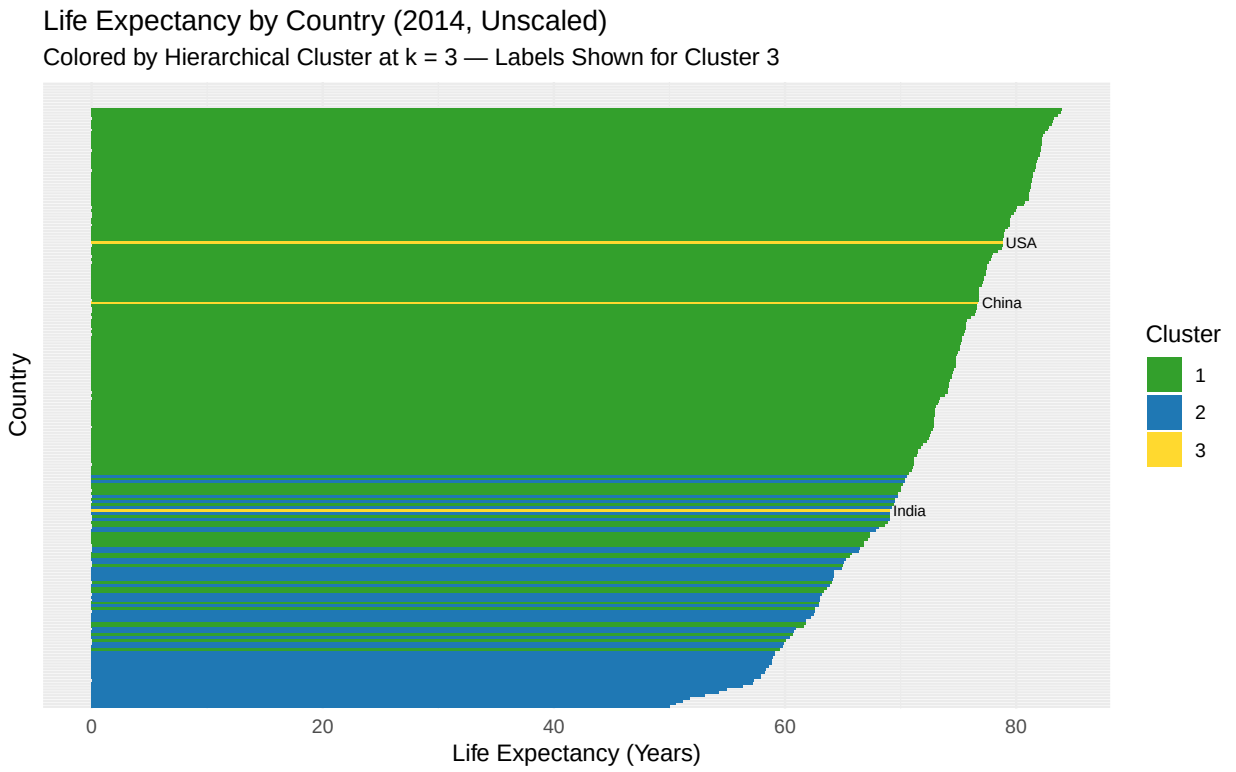


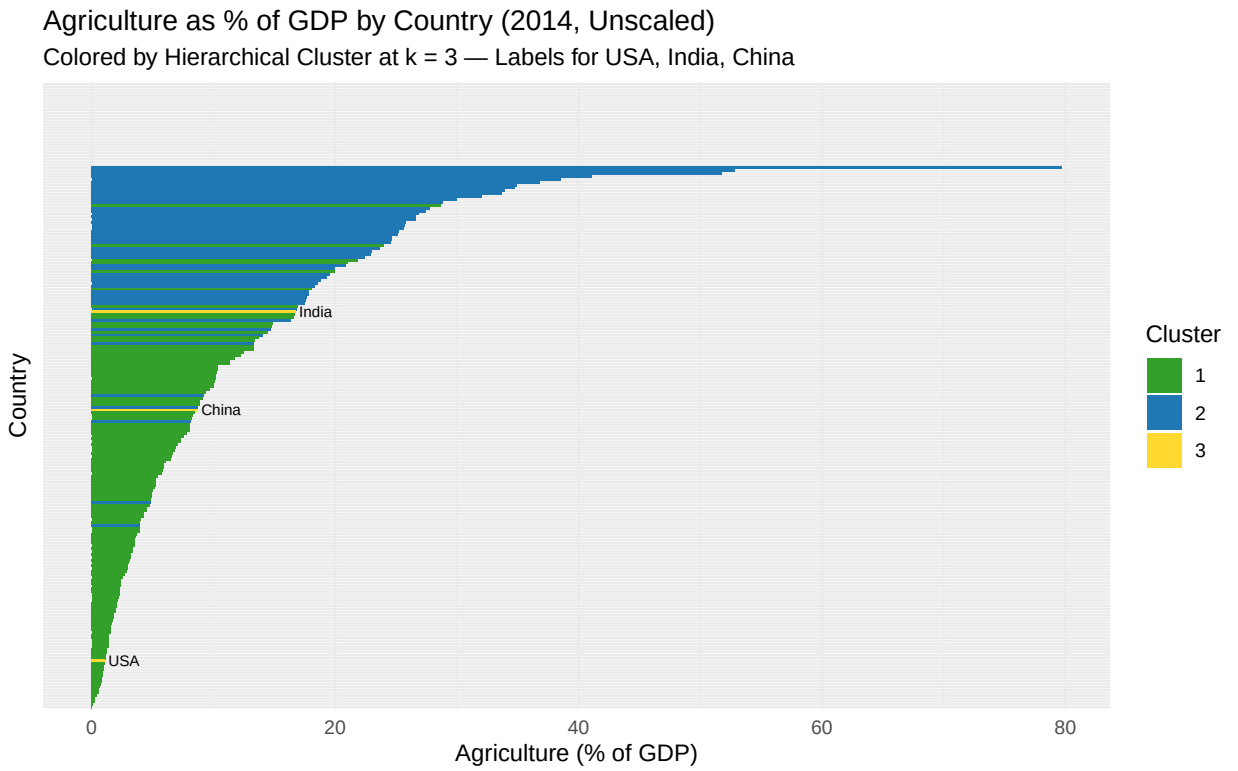
Interestingly, this reveals a cluster containing just three countries: China, India, and the United States. As k increases, this cluster persists until $k = 7$, where the cluster is split into a cluster containing China and India,

and a cluster containing the United States. The HDI of the countries align with the "developed" cluster, however further analysis indicates that this cluster has significant outliers across a number of indicators. This suggests that despite similar development levels by some metrics, their broader economic and social profiles are distinct enough to warrant separate clusters at higher resolutions.




Random Forest Variable Importance
Predicting Hierarchical Clusters (k = 3) with Scaled Data
AgricultureValueofGDP
LifeExpectancy
UrbanPop
EnergyUse
InternetEst
Exports
Revenue
Population
FinalConsumptionExpenditure
GDP
GrossSaving
AgriculturalArea
HealthcareExpenditure
Feature
0
5
10
15
Feature Importance (Gini Impurity Reduction)

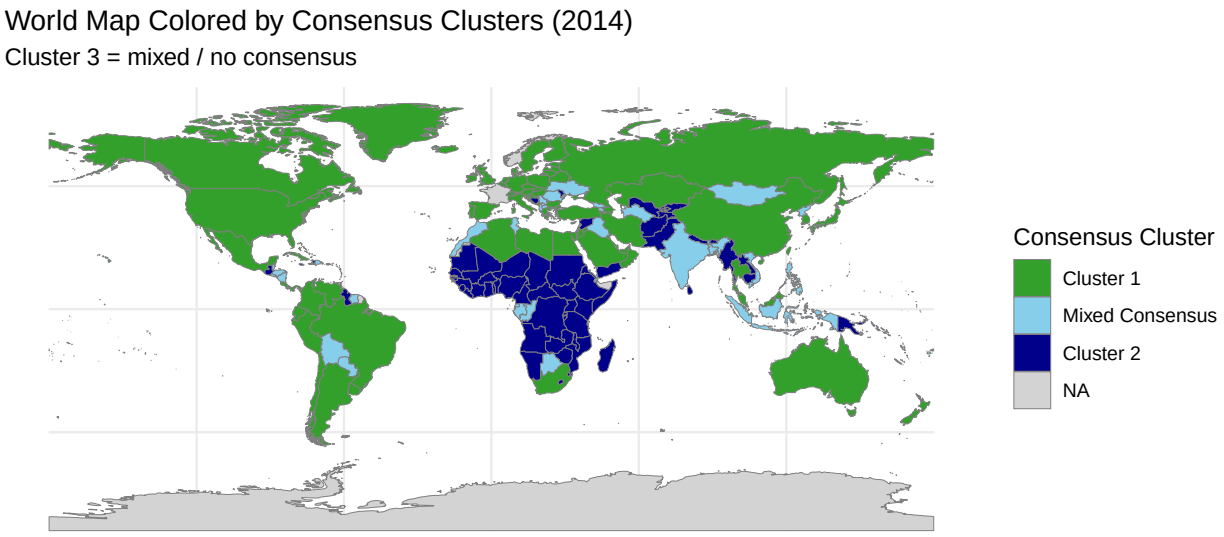

Life Expectancy by Country (2014, Unscaled)
Colored by Hierarchical Cluster at k = 3 — Labels Shown for Cluster 3
USA
China
India
Country
Cluster
1
2
3
0
20
40
60
80
Life Expectancy (Years)

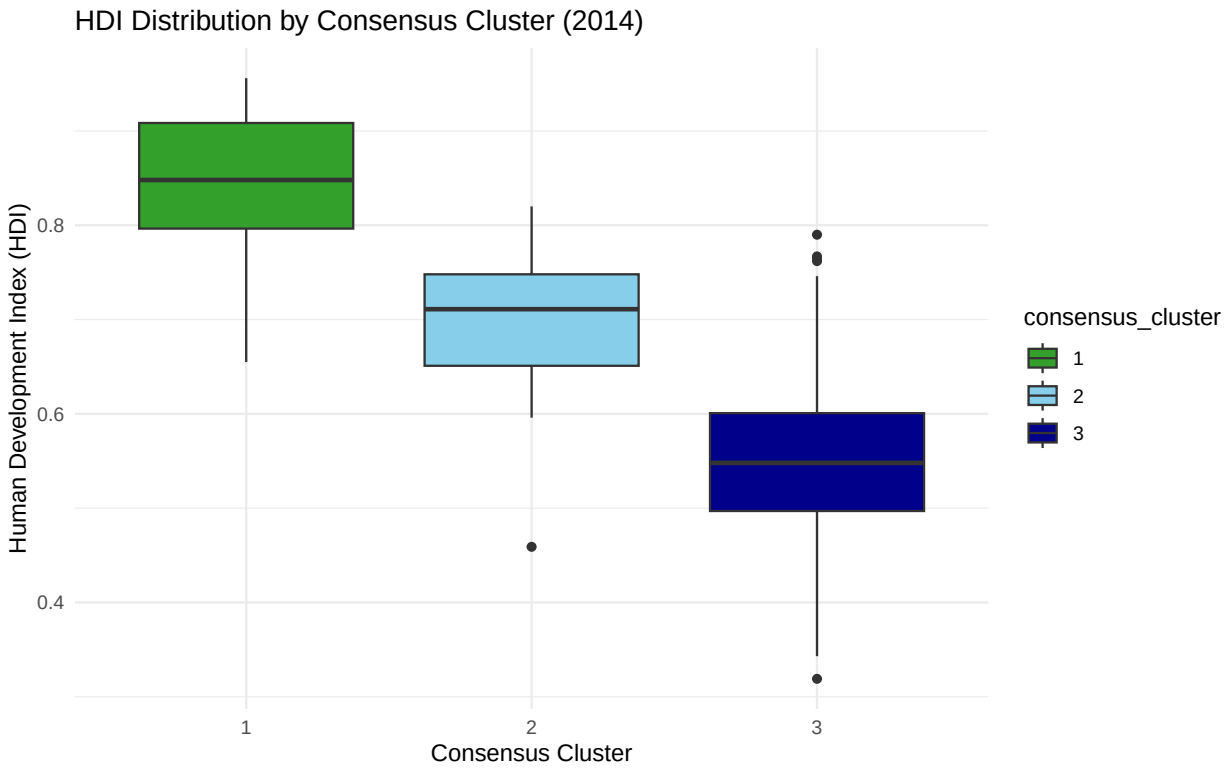


Notable outlying indicators for this new cluster include Population, GDP, and Exports. However, when using a random forest model and indicator visualizations, this cluster does not stand out on the most influential variables (outliers are highlighted in red on the importance plot). The extreme values in GDP, population, and export volume likely explain why these countries are particularly challenging for clustering algorithms to classify confidently. This also helps clarify why earlier methods showed uncertainty in the placement of countries like India. These plots highlight a strong positive correlation between life expectancy and HDI, and a strong negative correlation between agriculture's share of GDP and HDI. The former is expected, as life expectancy is one of the three main components used when generating the HDI value, however, the latter reflects the trend between economic diversification and higher human development. Notably, the United States stands out with the highest life expectancy and the lowest percentage of GDP from agriculture, which helps explain why it split from the third cluster before India or China.

Having examined each clustering method individually, I will now shift focus to how they align collectively. To synthesize results across methods, two approaches were used. First, a majority consensus approach was applied, where countries were grouped based on whether at least three out of four methods assigned them to the same cluster. Countries with two methods in each cluster were labeled as having mixed consensus. This formed a new three-level classification. In the second approach, I tallied the number of times each country was assigned to Cluster 1 across the four methods. This count—ranging from 0 to 4—was used

to form a new five-level classification. This allows us to assess how repeated cluster assignments relate to HDI, with more frequent Cluster 1 assignments typically corresponding to higher development. Together, these approaches provide a broader, more robust view of global development patterns through the lens of unsupervised clustering.



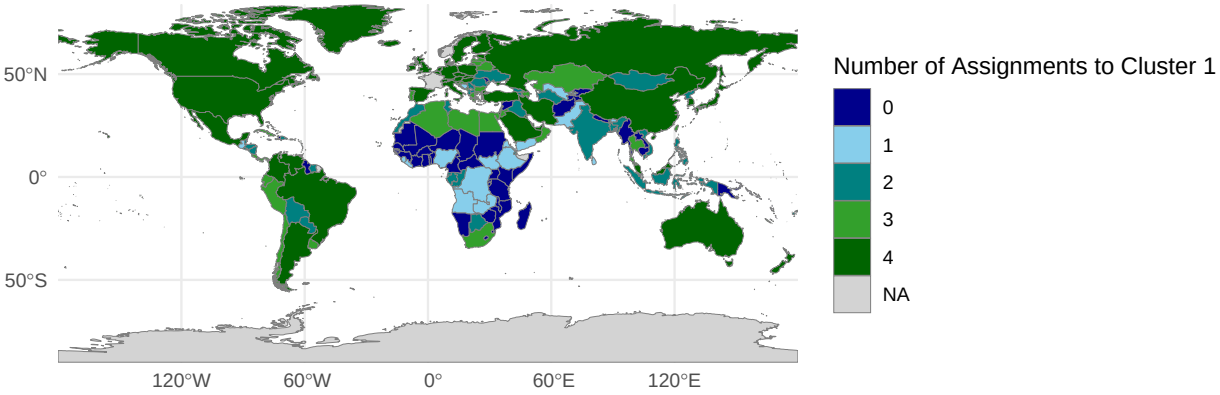


The consensus based approach produced a coherent global pattern: the Cluster 1 consensus group exhibited the highest average HDI, followed by the mixed group, with the Cluster 2 consensus group having the lowest. This suggests that clustering agreement is positively correlated with development level. Notably, the four South American countries with the lowest HDI are classified as either Mixed Consensus—such as Bolivia, Paraguay, and Suriname—or Cluster 2, which includes Guyana, the country with the lowest HDI in South America. This pattern extends to Latin America, where Haiti and Guatemala—the first and third lowest HDI countries in North America—are assigned to the Cluster 2 consensus group, while Honduras, with the second lowest HDI, falls into the Mixed Consensus group. In our second approach, I count the number of Cluster 1 assignments to refine the relative development ranking by introducing more nuanced cluster categories.

Next, I created an HDI map based on new cluster categories derived from the count of Cluster 1 assignments. Specifically, I defined HDI groups by the mean HDI values within each cluster count category (0 to 4 assignments). This approach allows us to visualize how well the clustering consensus corresponds to actual development levels, highlighting any inaccuracies or inconsistencies arising from the varying reliability of clustering methods. In other words, for each count of Cluster 1 assignments, I calculated the mean HDI and use these values as cutoffs to group countries into HDI ranges. I then mapped these groups to better understand how the number of Cluster 1 assignments relates to human development across countries.

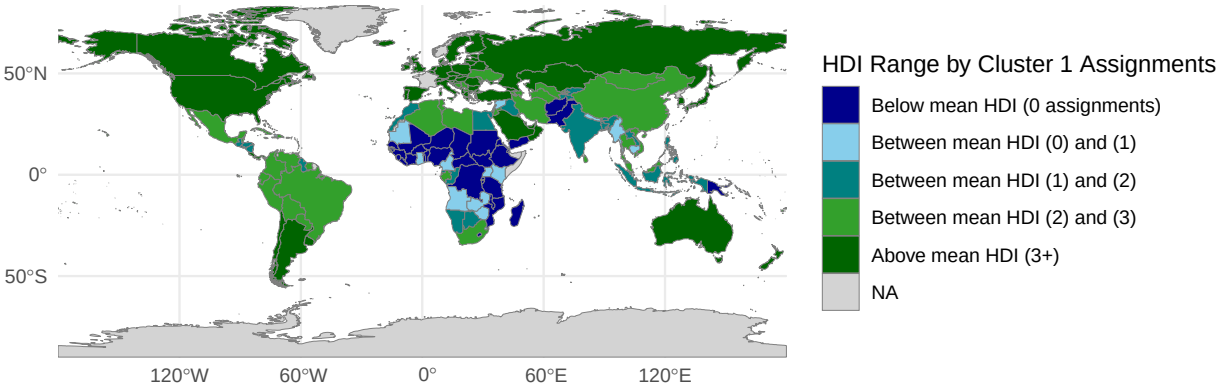
World Map by Number of Cluster 1 Assignments (2014)

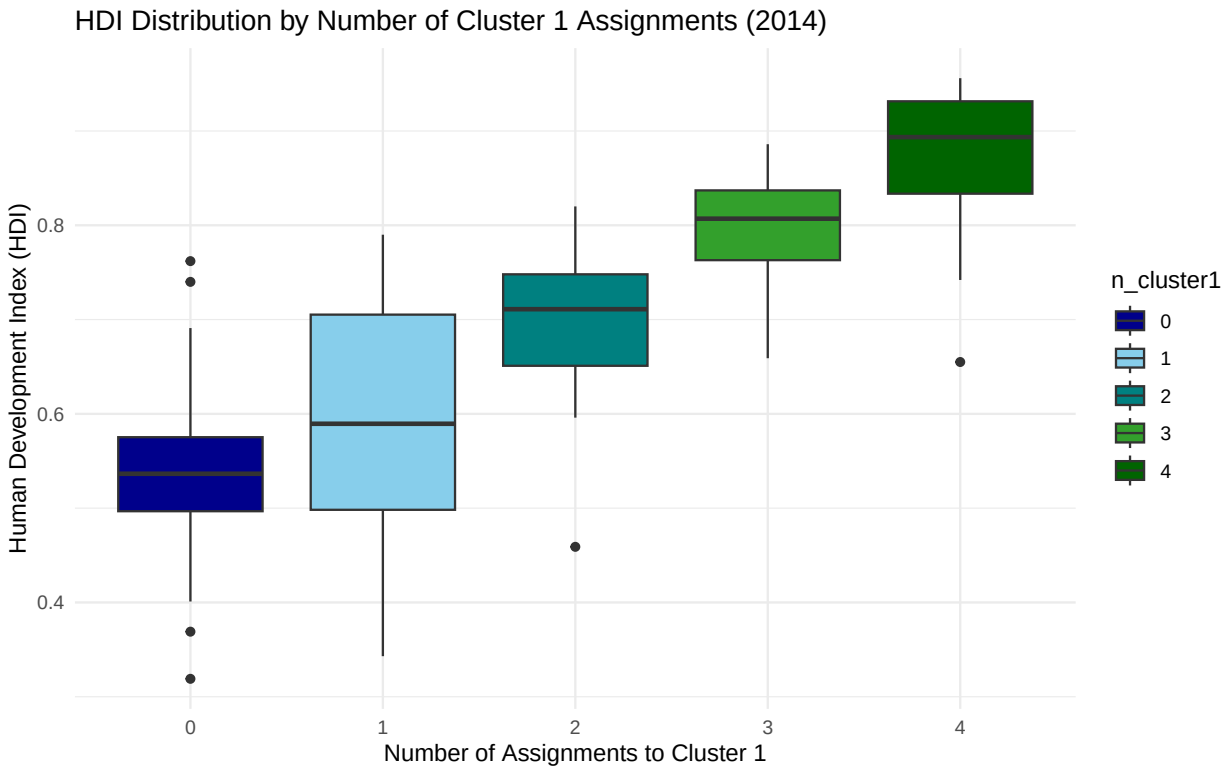
Color intensity reflects number of clustering methods assigning Cluster 1



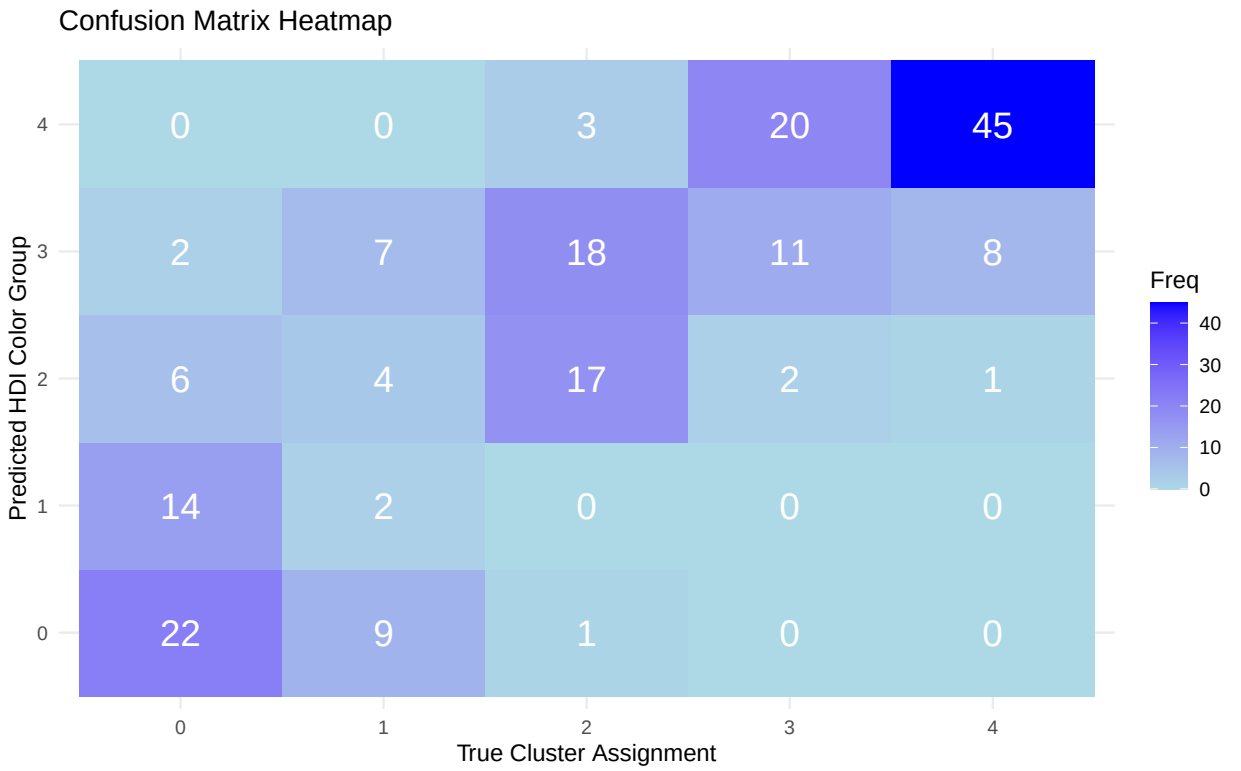
World Map Colored by HDI Range Relative to Cluster 1 Assignments

Colored by HDI intervals defined by mean HDI for each count of cluster 1 assignments





This approach highlights inaccuracies in our counting method predictions. To name a few, the counting method as predicted a wider variability in development within South America than HDI scores indicate. Additionally the inaccuracy of methods like GMM means our predictions for some developing countries such as the DROTC, South Sudan, and Pakistan are overestimated. However overall it seems that this provides a relatively accurate approximation of HDI. The bar plot shows a similar increasing HDI trend across groups as in our first approach, with some variation in countries assigned to Cluster 1 by only one method—likely due to GMM's lower accuracy. A confusion matrix further helps assess prediction accuracy.



This confusion matrix reveals that, although some inaccuracies occur—especially between similar clusters like 3 versus 4 assignments to Cluster 1—the approach is largely accurate when distinguishing developed from developing countries. These "one-off" misclassifications are likely due to limitations in the GMM model or other specific model inaccuracies.

Overall, the clustering approaches demonstrate a consistent ability to differentiate development levels, with some expected variation stemming from model-specific limitations. Having established these patterns and their associated uncertainties, we now turn to a deeper discussion of the implications, potential improvements, and broader context of these findings.

Discussion

Our analysis shows that while clustering methods consistently capture broad patterns in global development, there remain notable inaccuracies and limitations, particularly with methods like Gaussian Mixture Models (GMM). GMM's probabilistic framework appears sensitive to assumptions about data distribution, often overestimating development for central African countries. In contrast, Hierarchical Clustering demonstrated greater alignment with expected development patterns and produced clusters which reflected human development more closely. To improve our model and clustering reliability, several avenues could be explored:

1. **Incorporation of additional indicators**: Expanding the feature set to include more socio-economic indicators, such as governance metrics or inequality indices, may enhance cluster differentiation.

2. **Iterative or Ensemble Hierarchical Clustering**: By running hierarchical clustering multiple times with varied seeds or using different linkage criteria (e.g., average, complete, Ward's) could generate diverse cluster solutions. Aggregating these via consensus clustering or co-association matrices could stabilize cluster assignments and identify robust groupings.

3. **Exploring Alternative Clustering Algorithms**: Incorporating additional clustering methods such as density based methods such as DBSCAN or HDBSCAN may help identify atypical countries that do not conform to standard development patterns.

4. **Addressing Missing Data and Imputation Effects**: Some countries, such as France and Norway, were excluded due to extensive missing data. Exploring alternative imputation methods beyond K-Nearest Neighbors (KNN) may offer better handling of incomplete records and reduce bias in clustering outcomes. Assessing how these choices affect cluster assignments could improve the robustness of the overall model.

Conclusion

This project set out to explore global patterns of development using unsupervised clustering methods, with the goal of reinterpreting the traditional "First World" vs. "Third World" dichotomy through modern economic and social indicators. By applying and comparing several clustering techniques—K-means, UMAP, Gaussian Mixture Models (GMM), and Hierarchical Clustering—we examined how different algorithms interpret global disparities in development and which produce the most reasonable groupings relative to Human Development Index (HDI) scores.

Hierarchical Clustering emerged as the most accurate and interpretable method, producing results that aligned closely with HDI and known regional development levels. UMAP also performed well, particularly in preserving nuanced regional distinctions, while K-means occasionally overgeneralized. GMM proved less reliable, frequently misclassifying countries due to its sensitivity to distributional assumptions which are not well-suited to the data.

An ensemble approach using multiple techniques proved to provide accurate classifications, often identifying regions more accurately than any one technique. Despite the model's strengths, several limitations remain. Key countries were excluded due to missing data, and the choice of imputation method (KNN) may have influenced results. Additionally, the project used a limited set of indicators, which, while illustrative, do not capture the full range of development dimensions. Future work could expand the model by incorporating more variables and by exploring iterative or hybrid clustering methods.

While clustering cannot replace human judgment or institutional definitions of development, it offers a powerful tool for uncovering patterns, identifying outliers, and challenging simplistic categorizations. Through careful method selection, consensus-building, and continual refinement, unsupervised learning can help uncover the intricate structure of global development disparities.

Data Sources

The following indicators and datasets were used in this project:

1. **Exports of goods and services (current US$)**. World Bank.
   https://data.worldbank.org/indicator/NE.EXP.GNFS.CD
2. **GDP (current US$)**. World Bank.
   https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
3. **Final consumption expenditure (% of GDP)**. World Bank.
   https://data360.worldbank.org/en/indicator/WB_WDI_NE_CON_TOTL_ZS
4. **Agriculture, value added (% of GDP)**. World Bank.
   https://data360.worldbank.org/en/indicator/FAO_AS_4113
5. **Gross savings (% of GDP)**. World Bank.
   https://data360.worldbank.org/en/indicator/WB_WDI_NY_GNS_ICTR_ZS
6. **Agricultural land (% of land area)**. World Bank.
   https://data.worldbank.org/indicator/AG.LND.AGRI.ZS
7. **Population, total**. World Bank.
   https://data.worldbank.org/indicator/SP.POP.TOTL
8. **Revenue (% of GDP)**. International Monetary Fund / World Bank.
   https://data360.worldbank.org/en/indicator/IMF_FM_GGR_G01_GDP_PT
9. **Life expectancy at birth, total (years)**. World Bank.
   https://data360.worldbank.org/en/indicator/WB_WDI_SP_DYN_LE00_IN
10. **Urban population (% of total population)**. World Bank.
    https://data360.worldbank.org/en/indicator/WB_WDI_SP_URB_TOTL_IN_ZS
11. **Current health expenditure (% of GDP)**. World Bank.
    https://data360.worldbank.org/en/indicator/WB_WDI_SH_XPD_CHEX_GD_ZS
12. **Energy use (kg of oil equivalent per capita)**. World Bank.
    https://data360.worldbank.org/en/indicator/WB_WDI_EG_USE_PCAP_KG_OE

13. **Secure Internet servers (per 1 million people)**. World Bank.
    https://data.worldbank.org/indicator/IT.NET.SECR.P6

14. **Human Development Index (value)**. United Nations Development Programme (UNDP).
    https://hdr.undp.org/data-center/documentation-and-downloads