Code

End-to-End Economic & Social Indicators Analysis

How Economic, Health, and Education Factors Drive Human Development

Sana Ur Rehman Arain

2026-01-26

---

1. Executive Summary

This project investigates the **primary drivers of the Human Development Index (HDI)** using **4,200+ country–year observations (2000–2022)**. By integrating **World Bank economic indicators** with **UNDP human development data**, the analysis moves beyond correlation toward **predictive and explanatory modeling**.

Key Findings

- **Non-Linearity Matters:** Random Forest outperformed Linear Regression by **54%** (RMSE **0.026 vs 0.056**), confirming that development follows complex, non-linear patterns.
- **Wealth vs. Well-Being:** While GDP per capita is the strongest single predictor, **Health Expenditure** and **Unemployment Rate** act as critical *bottlenecks* to development.
- **The Preston Curve Validated:** Economic growth delivers rapid HDI gains for low-income countries but **plateaus** among high-income economies.

---

2. Data Engineering & Quality

Two complementary datasets were combined using a rigorous **ETL (Extract–Transform–Load)** pipeline.

Data Sources

- **World Bank API:** Economic, labor, education, and health indicators
- **UNDP:** Human Development Index (HDI)

Key Processing Steps

- Harmonized country identifiers using **ISO3C codes**
- Constructed a balanced panel (2000–2022)
- Imputed missing values using **group-wise Down–Up filling**

Final Dataset

- **4,290 observations**
- **15 development indicators**

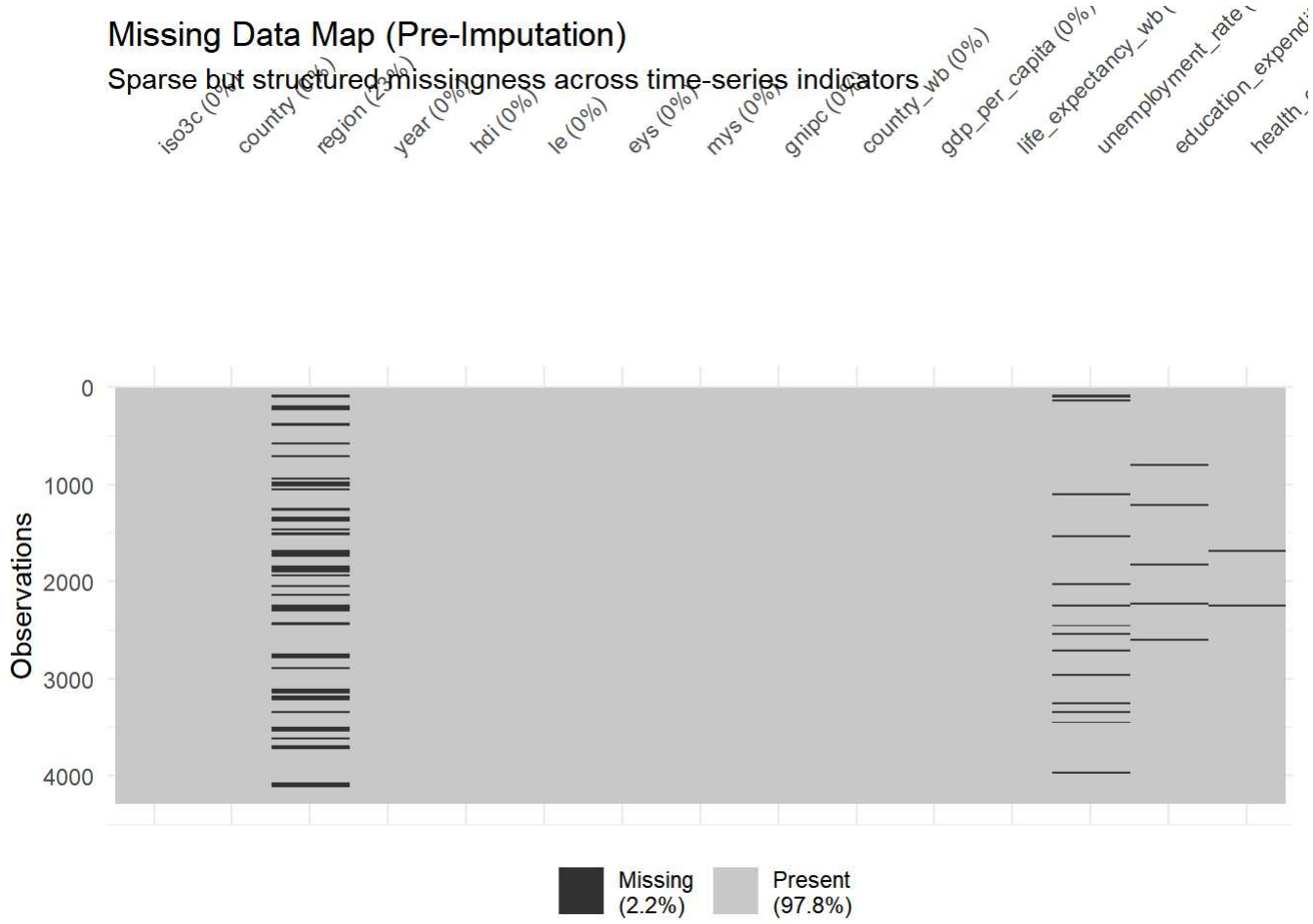
Show

country	year	hdi	gdp_per_capita	life_expectancy_wb
Czechia	2006	0.864	16107.3043	76.52439
Niger	2015	0.374	485.6626	58.85000
Benin	2007	0.469	926.7401	58.24900
Marshall Islands	2012	0.668	3420.9144	64.77600
Mauritius	2012	0.778	8702.8224	73.86341

3. Missing Data Diagnostics

Before imputation, we examined the structure of missingness to ensure data integrity.

Show

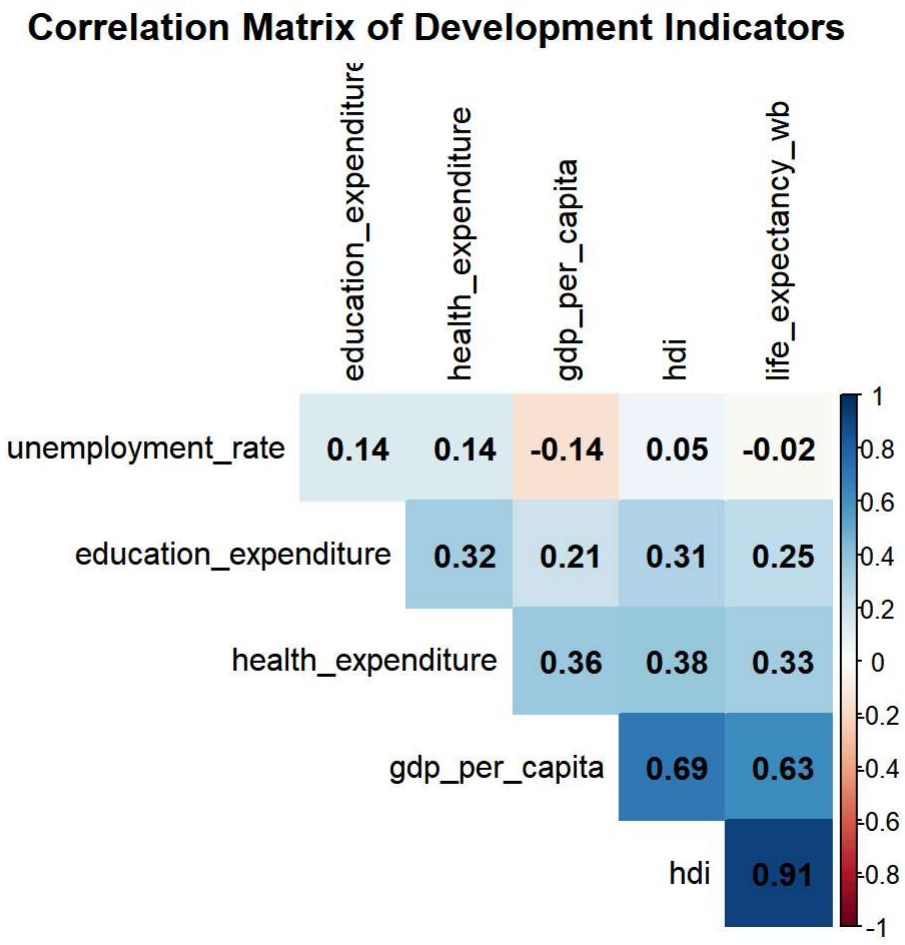


4. Exploratory Data Analysis (EDA)

4.1 Correlation Structure

Strong linear relationships emerge between HDI and health-related variables, particularly **Life Expectancy**.

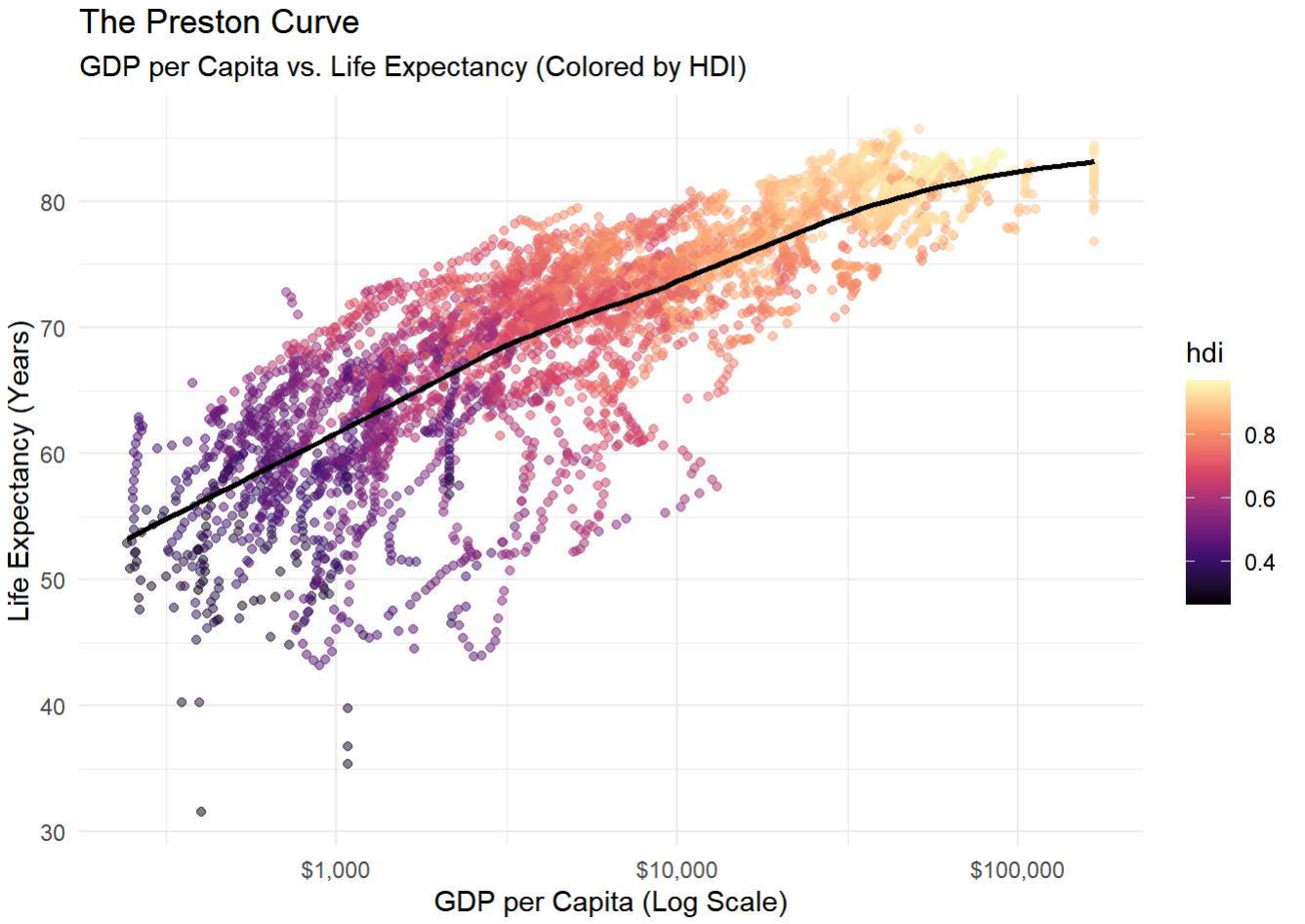
Show



4.2 The Preston Curve (Wealth vs. Health)

This figure visualizes **diminishing returns of income on health outcomes**.

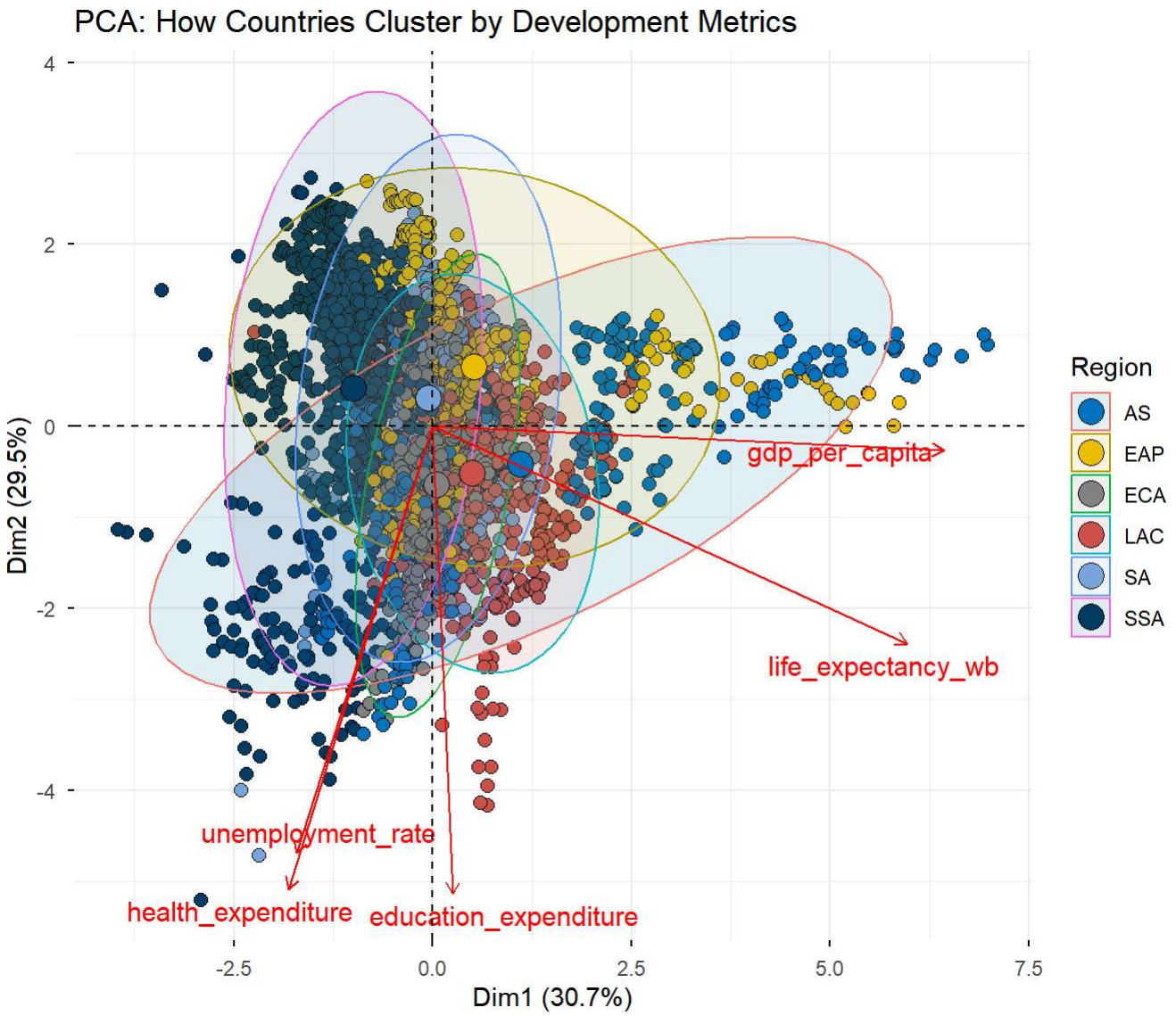
Show



4.3 Structural Clustering via PCA

Principal Component Analysis reveals **distinct regional development patterns**.

Show



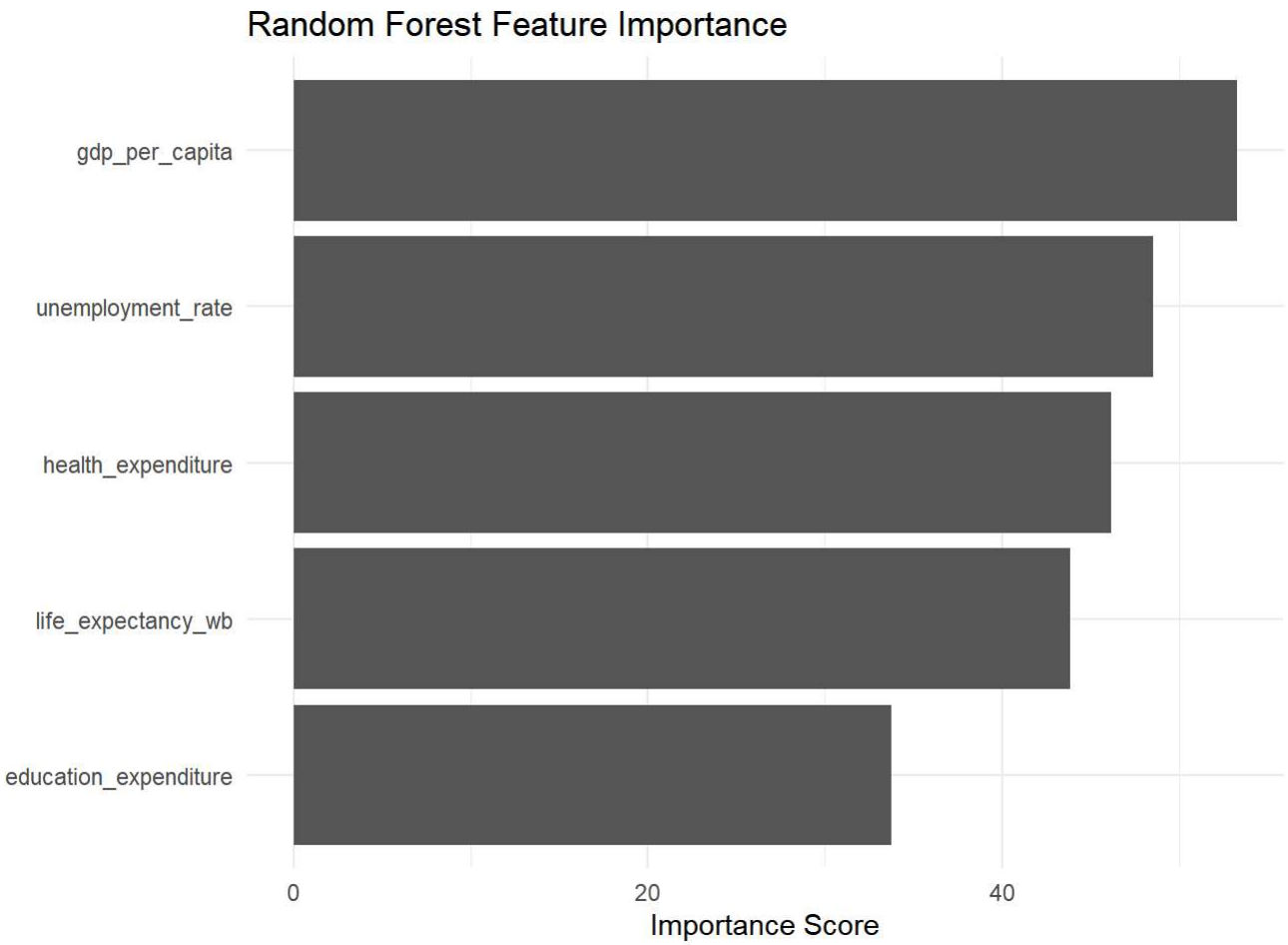
5. Statistical Modeling & Machine Learning

We compare **Linear Regression** and **Random Forest** using an **80/20 train–test split**.

Show

5.1 Feature Importance (Random Forest)

Show



5.2 Model Performance Comparison

Show

Model	RMSE
Random Forest	0.0259180
Linear Regression	0.0562847

6. Conclusion

This analysis demonstrates that **economic growth is necessary but not sufficient** for human development. Once basic income thresholds are reached, **health investment and labor market conditions** become decisive.

From a policy perspective, the results strongly suggest that **social infrastructure—not income alone—drives long-run human well-being**.