Socio-Economic Country Profiling Using a Lakehouse Architecture and K-Means Clustering

A Data-Driven Analysis Based on the World Values Survey

Achraf IKISSE, Ammar KASBAOUI, Ilias ISSAF
Université Internationale de Rabat

Academic Year 2025–2026

Abstract

Measuring socio-economic well-being across countries is a complex and multidimensional problem that cannot be captured by economic indicators alone. This paper proposes a data-driven approach combining a Lakehouse architecture with unsupervised machine learning to identify country profiles based on subjective and social indicators extracted from the World Values Survey (WVS). Using K-Means clustering, countries are grouped into distinct socio-economic profiles. The results highlight meaningful patterns in well-being and provide actionable insights for public policy and comparative social analysis.

1 Introduction and Problematic

Evaluating societal well-being is a major challenge for governments and international organizations. Traditional economic indicators such as GDP per capita provide an incomplete view of social reality, as they fail to capture subjective dimensions like happiness, perceived health, education satisfaction, and employment stability.

Research question: *How can countries be objectively grouped according to multidimensional well-being indicators derived from large-scale survey data?*

To answer this question, we design a complete analytical pipeline relying on a Lakehouse data architecture and unsupervised clustering.

2 Data and Methodology

2.1 Dataset Description

The World Values Survey (Wave 7) contains more than 97,000 individual responses across 66 countries. Five variables were selected due to their relevance to socio-economic well-being: country, happiness, health perception, education level, and employment status.

2.2 Lakehouse Architecture

The data pipeline follows the Bronze–Silver–Gold paradigm:

- **Bronze**: Raw CSV survey data
- **Silver**: Cleaned individual-level dataset
- **Gold**: Aggregated country-level indicators

This architecture ensures data quality, reproducibility, and analytical scalability.

2.3 Feature Engineering

Each country is represented by four aggregated indicators:

- Mean happiness
- Mean perceived health
- Mean education level
- Employment diversity (number of distinct employment categories)

2.4 Clustering Model

K-Means clustering is applied with $k = 3$ after feature standardization.

Listing 1: K-Means clustering

```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

X_scaled = StandardScaler().fit_transform(X)
kmeans = KMeans(n_clusters=3, random_state=42)
labels = kmeans.fit_predict(X_scaled)
```

3 Statistical Analysis

Figure 1 shows the distribution of happiness across individuals, highlighting strong variability.

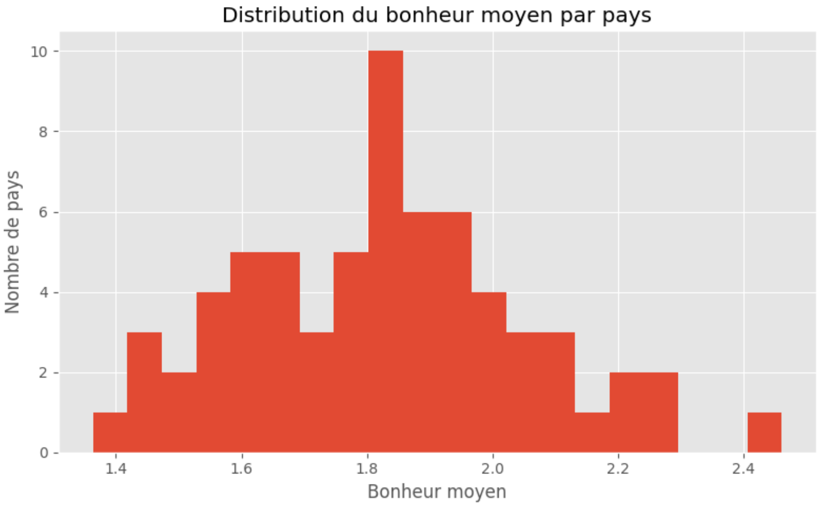

Figure 1: Distribution of happiness levels across individuals

Figure 2 presents the ten countries with the highest average happiness.

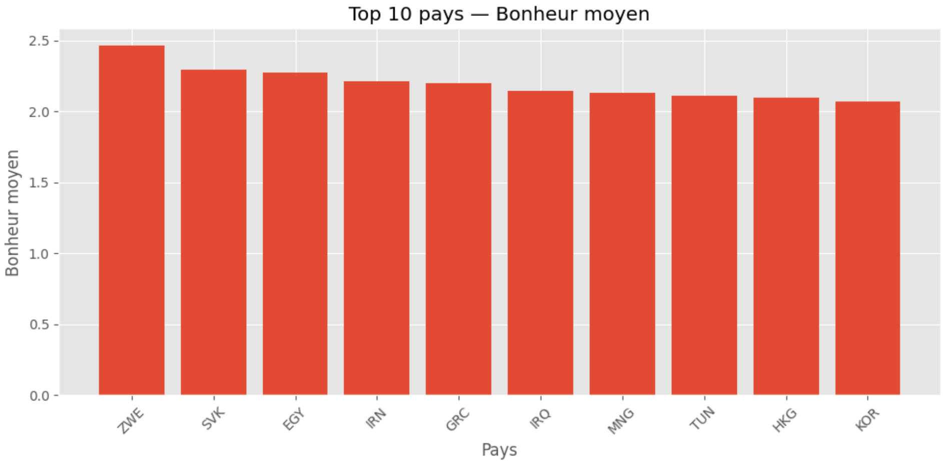

Figure 2: Top 10 countries by mean happiness

4 Clustering Results

The K-Means algorithm identifies three distinct country clusters. Figure 3 illustrates the cluster separation, while Figure 4 shows the number of countries in each cluster.

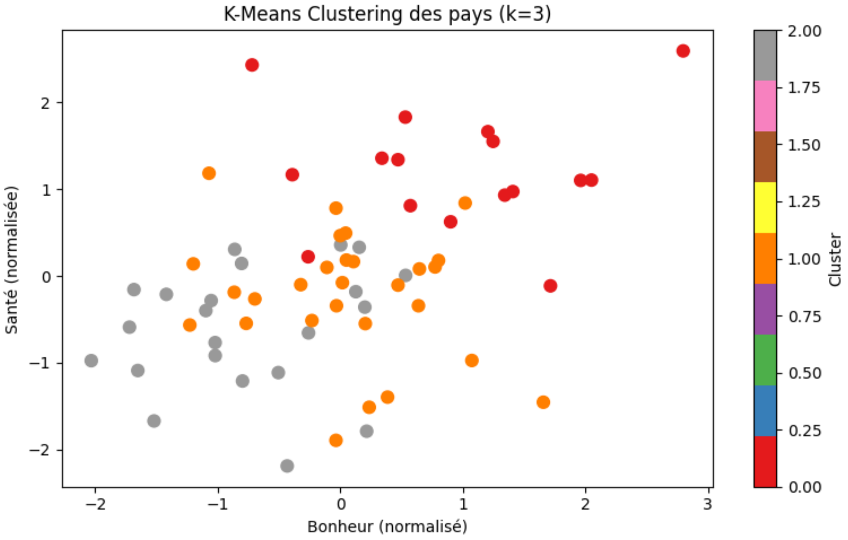

Figure 3: K-Means clustering of countries (standardized space)

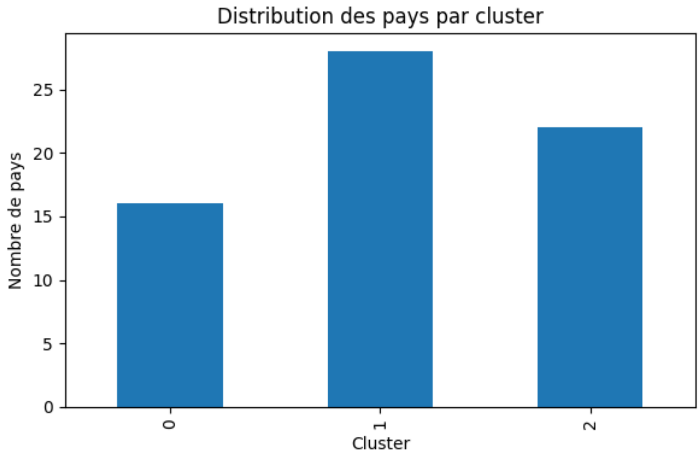

Figure 4: Distribution of countries across clusters

The clusters can be interpreted as follows:

- **Cluster 0**: High happiness and strong health indicators
- **Cluster 1**: Intermediate socio-economic well-being
- **Cluster 2**: High education but lower happiness levels

4.1 Representative Countries by Cluster

To improve interpretability, Table 1 presents representative countries from each cluster.

Table 1: Representative countries by cluster

Cluster	Profile Interpretation	Example Countries
Cluster 0	High happiness and health	Australia, Greece, Slovakia, Zimbabwe
Cluster 1	Intermediate well-being	Argentina, Egypt, Iran, Bangladesh
Cluster 2	High education, lower happiness	Armenia, Uzbekistan, Kyrgyzstan, Tajikistan

5 Discussion

The clustering confirms that socio-economic well-being is inherently multidimensional. While education is often associated with positive outcomes, the results show that higher education levels do not necessarily translate into higher happiness. Social stability, health perception, and employment structure play a crucial role in shaping subjective well-being.

6 Recommendations

Based on the analysis, we recommend:

- Incorporating subjective well-being indicators into national policy evaluation frameworks
- Prioritizing health and social cohesion policies in lower-happiness clusters
- Using cluster-based benchmarking to design targeted socio-economic interventions

7 Conclusion

This study demonstrates the effectiveness of combining a Lakehouse architecture with unsupervised learning for large-scale socio-economic analysis. The proposed pipeline enables interpretable country profiling and supports data-driven decision-making. Future work may extend this framework to temporal analysis across multiple survey waves or alternative clustering techniques.