Colorado State University

Mapping Subjective Social Status Through Data and Models

Final Project

Neha Deshpande, Witlie Leslie, Phuong Nguyen

DSCI 478

Dr. King

Github Repository: https://github.com/witlie/child-social-status

Introduction

Social class has a well-known impact on health, but it is usually measured through things like income, education, or job status. While these factors are important, they don't always tell the full story. This study, *Measuring subjective social status in children of diverse societies,* looks at whether subjective social status can also predict how healthy they feel. Subjective social status is how people perceive themselves in social hierarchy. Using survey data from four countries, the researchers test whether people's self-rated health is influenced by where they think they stand socially, even after accounting for their actual socioeconomic background. We picked this topic because we were interested in how kids see their own place in society, not just how much money their families make or what jobs their parents have. This project gave us a chance to explore that idea using real data from different countries and apply tools like data visualization and machine learning to better understand those patterns in a meaningful way.

Exploratory Analysis

A critical part of data analysis is doing some exploratory work on the original dataset. The dataset includes information on children's self-perceived social status and their access to basic material needs. The main variable of interest is stairs, which represents where a child places themselves on a social ladder, indicating their perceived position in society. Several demographic and socioeconomic factors are used as predictors. The dataset used in the study included a diverse range of countries, allowing for analysis across different geographic regions.

The exploratory data analysis focused on understanding the relationship between children's material conditions and their perceived social status. This portion of the EDA was focused on India and Argentina. Correlation analyses were conducted between the perceived social status variable (stairs) and four material well-being indicators: Qhouse, Qfood, Qmoney, and Qthings. Results indicated that in India, access to money and personal items had the strongest positive correlations with higher social status, whereas in Argentina, access to food showed the strongest association.

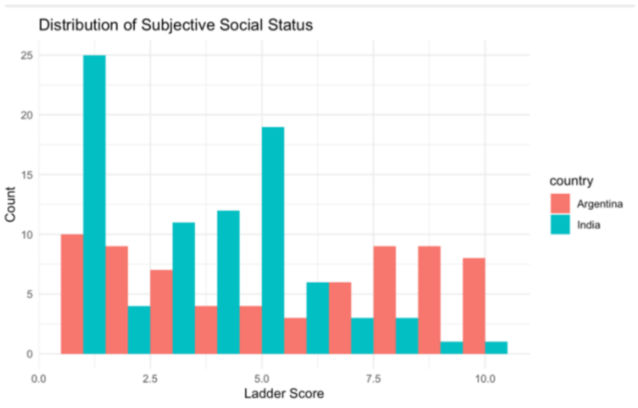


Figure 1. Distribution of Subjective Social Status for Argentina and India.

This next portion will explore the socioeconomic perceptions among children from two rural regions in Ecuador: Cross Cutucú and Upano Valley. Using survey data from the dataset, we compared responses to four “ladder” questions—housing quality, food access, financial stability, and material possessions—rated on a 0 to 3 scale. These self-reported scores provide insight into how children perceive their everyday living conditions. The demographic profiles of the two groups are quite similar. The average age in Cross Cutucú is 10.8 years, while in Upano Valley it is 10.6 years. Gender distribution is also nearly identical, with approximately 53% of participants identifying as male in Cross Cutucú and 52% in Upano Valley. This similarity in demographic makeup allows for a meaningful comparison of perceived socioeconomic status between the two regions.

Across all four ladder domains, children from Cross Cutucú consistently reported slightly higher scores. The average housing score was 1.79 in Cross Cutucú compared to 1.70 in Upano Valley. Food access stood out most sharply, with Cross Cutucú averaging 2.18 versus 1.94 in

Upano Valley. Perceived financial status (Qmoney) also followed this trend, with Cross Cutucú scoring 2.34 and Upano Valley 2.11. Material possessions (Qthings) were rated at 2.00 and 1.96 respectively.

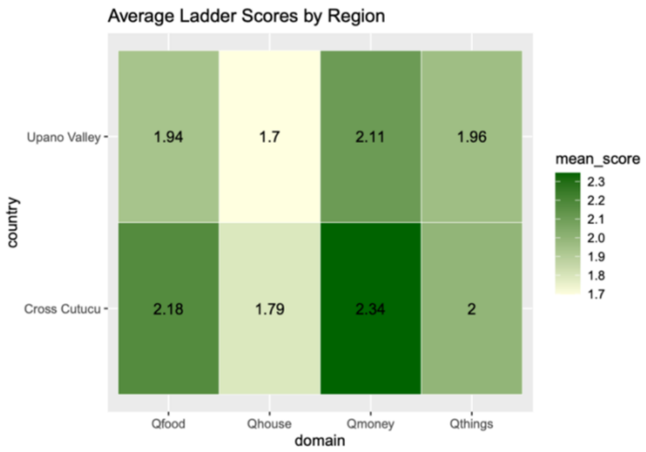


Figure 2. Average Ladder Scores from Upano Valley and Cross Cutucu.

These findings suggest that children in Cross Cutucú may perceive greater stability or support in their immediate environment, especially regarding access to food and financial resources. This aligns with existing literature describing the role of traditional ecological knowledge and strong kinship ties in Shuar communities, which help buffer against economic instability and food insecurity (Open Case Studies, 2020). In contrast, Upano Valley, while geographically closer to urban development, has experienced more agricultural expansion and external market pressures, potentially contributing to less consistent access to basic resources (Ditmars, 2024). These differences in environmental and economic context may explain the modest disparities in perceived socioeconomic well-being between the two groups.

Across all countries, age was normally distributed with a slight right skew and a peak at 8 years old. Ages ranged from 4 years old to 18 years old. The distribution of sex was uniform, with a similar number of male and female respondents.

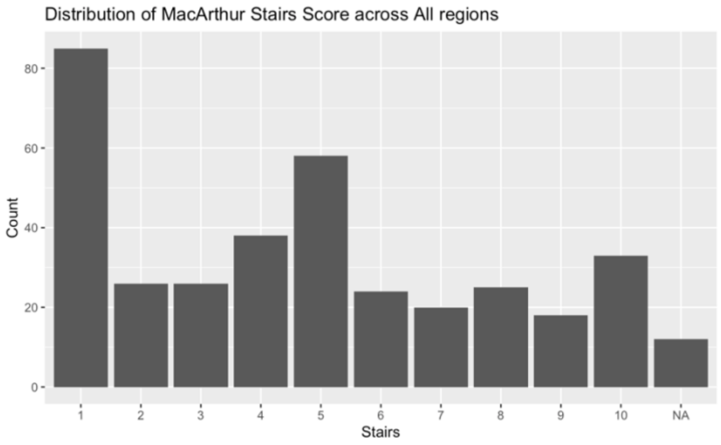


Figure 3. Bar Graph of Stair Responses Across All Participants.

Above is a histogram chart visualizing the distribution of stairs responses from all children in the dataset. This distribution is multimodal, with the most common response being 1. This indicates that many children feel as though their families are among those with the most money, schooling, and respect in their country. The second most common response was 5, indicating that these children feel as though their families are among the middle or average of their country in terms of money, schooling, and respect.

Cluster Analysis

To further investigate the relationships between variables, we performed a k means cluster analysis to identify groupings of similar responses. We hypothesized that perhaps we would find groupings of responses that shared interesting similarities, e.g., children who feel as though they have worse than their peers in terms of their house but don’t share this sentiment with food.

To determine the optimal number of clusters, we looked at the within-cluster sum of squares (WSS) for k values from 1-10 on an elbow plot. After k = 4, the decrease in WSS is less significant for each additional cluster, indicating that we should proceed forward using 4 clusters.

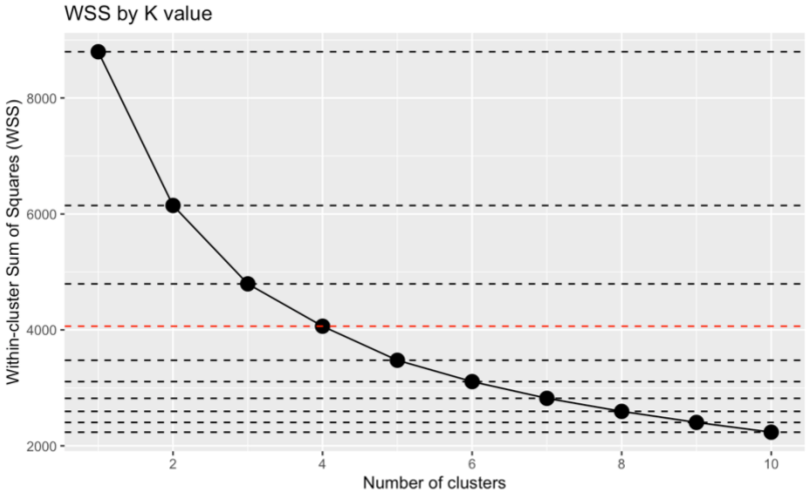


Figure 4. Within-cluster Sum of Squares by Number of Clusters.

We included all numerical variables from the dataset in our k means cluster analysis. Below is a visualization of the clustering results showing just two variables, age and stairs. We can see that our cluster analysis resulted in the following groupings: younger children with lower subjective social status, younger children with higher subjective social status, older children with lower subjective social status, and older children with higher subjective social status.

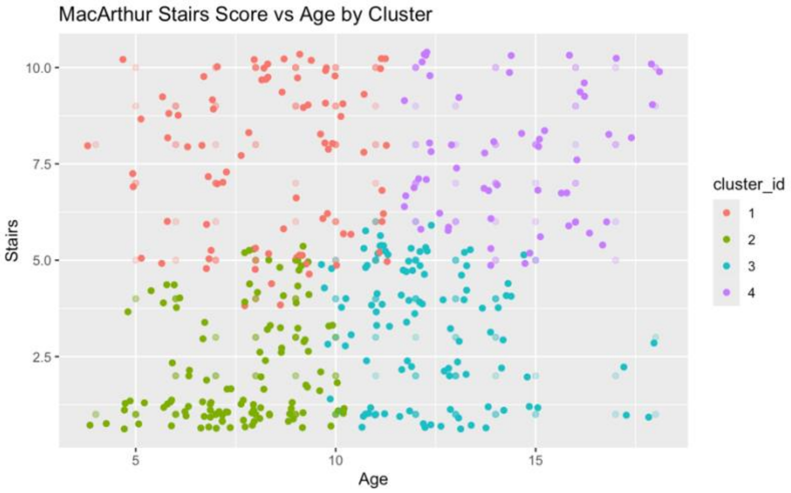


Figure 5. MacArthur Stair Value vs. Age by Cluster

Logistic Regression

To further investigate the relationships between variables in the dataset, we assessed multiple logistic regression models with varying interaction terms. We first tested the assumptions of linearity and minimal multicollinearity using a base model without interaction terms. The response variable is a binary term indicating either high social status on the MacArthur ladder (stairs score >= 5) or low social status (stairs score < 5). The lines shown below in the residual plots for each variable are approximately straight, which suggests linearity. The variance inflation factor (VIF) for each variable was < 2, indicating acceptable collinearity.

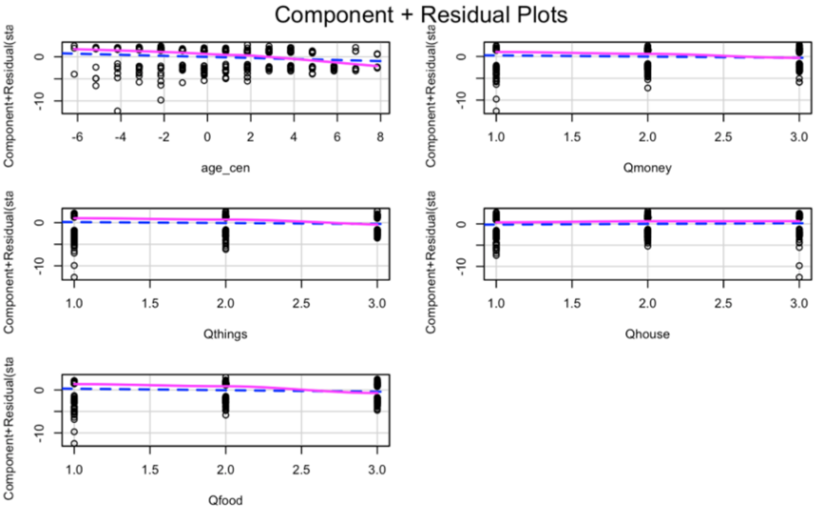


Figure 6. Residual Plots by Variable.

We tested five interaction terms: age and sex, country and Qmoney, age and Qthings, age and Qmoney, and sex and Qthings. The only interaction term that showed a significant correlation to the response variable binary social status was the interaction between country and Qmoney with a p-value of 0.012.

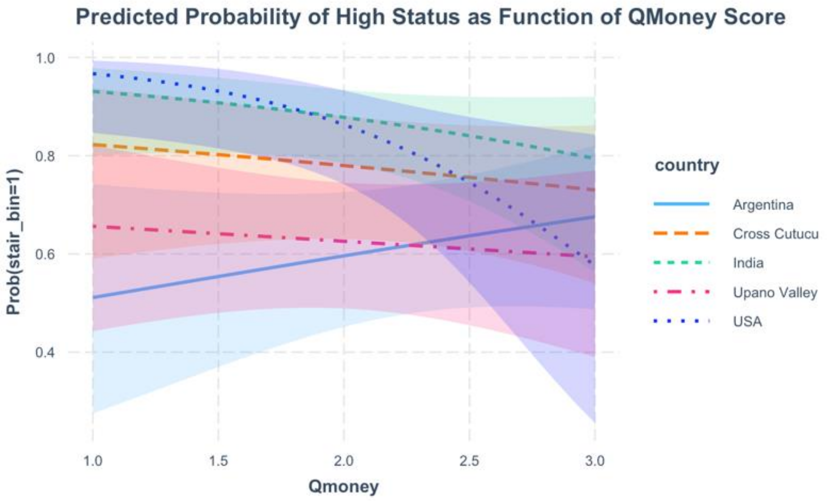


Figure 7. Predicted Probability of High Social Status as a Function of Qmoney by Country.

The graph above depicts the predicted probability that a child will assign themselves a high social status on the MacArthur ladder (stair >= 5) as a function of their comparative money score for each region. For most regions, probability of high status decreases as Qmoney score increases (the sense of having less money than others increases). This makes sense intuitively, as money is a significant factor in social status. Interestingly, this relationship is inverted for Argentina, with children who say they have less money than others being more likely to describe their family as having a high social status. This could suggest that money is less important to Argentinian children's perception of social status relative to other metrics (e.g., food, things, house).

Machine Learning Models

Random forest is a machine learning algorithm that's often used when you want to make predictions based on patterns in data. In our project, we used a random forest regression model to predict how many stairs someone can climb, based on their age, sex, country, and a new feature we created called *material_index* (an average of things like housing and food security.) The model only explained about 8.76% of the variance in the training data and 10% in the test set, meaning it wasn’t capturing much of the pattern behind the stair-climbing ability. The Mean Squared Error (MSE) on the test set was 6.92, which suggests our predictions weren’t very close to the real values.

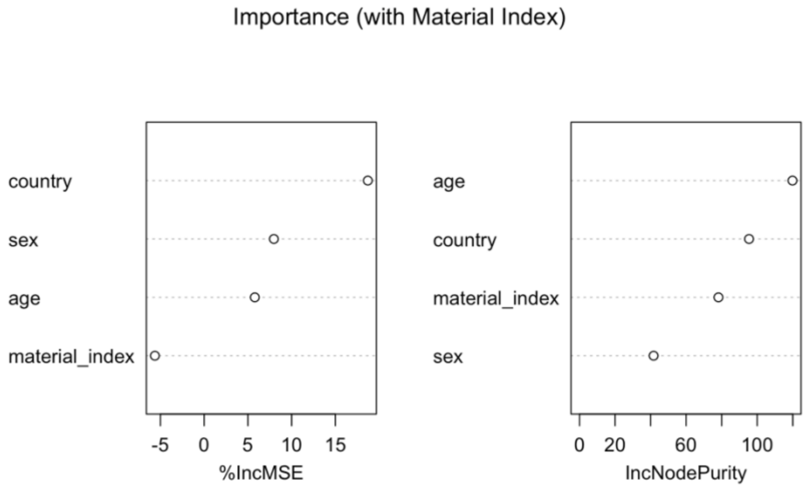


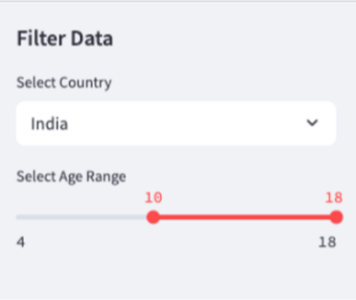
Figure 8. Variable Importance Plots.

We looked at two plots to measure importance: one based on how much the error increases when a variable is removed (%IncMSE), and another based on how often a variable is used to split the data (IncNodePurity). In the %IncMSE plot, material_index was the most important, meaning the model relies on it a lot to make accurate predictions. In the IncNodePurity plot, age was the most important, meaning it helps the model make better splits in the decision trees. Other variables like sex and country were less important in both plots. The model had a low error rate (8.31%) and a good F1 score (0.821), showing it works well overall.

Dashboard Description

To facilitate exploration and interpretation of subjective social status (SSS) across cultural contexts, we developed an interactive dashboard that integrates data visualization, machine learning models, and geographic mapping tools. The dashboard is organized into six main components:

1. Filter Panel



Filter Data

Select Country

India

Select Age Range

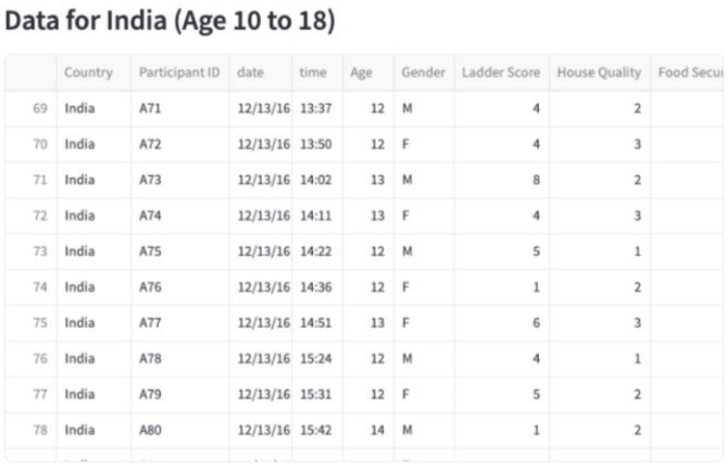
10 18

4 18

Figure 9. Filter Panel.

Located on the left sidebar; this panel allows users to filter the dataset by country and age range. In the example displayed, the data are filtered to include responses from children aged 10 to 18 in India. This interactive filtering ensures that all subsequent visualizations and model outputs are based on the specified subset of the data.

2. Filtered Data Table



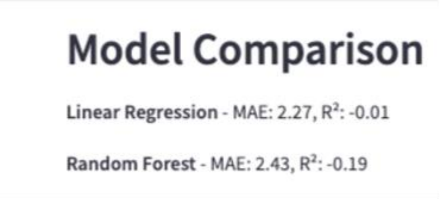
Data for India (Age 10 to 18)

	Country	Participant ID	date	time	Age	Gender	Ladder Score	House Quality	Food Secu
69	India	A71	12/13/16	13:37	12	M	4	2	
70	India	A72	12/13/16	13:50	12	F	4	3	
71	India	A73	12/13/16	14:02	13	M	8	2	
72	India	A74	12/13/16	14:11	13	F	4	3	
73	India	A75	12/13/16	14:22	12	M	5	1	
74	India	A76	12/13/16	14:36	12	F	1	2	
75	India	A77	12/13/16	14:51	13	F	6	3	
76	India	A78	12/13/16	15:24	12	M	4	1	
77	India	A79	12/13/16	15:31	12	F	5	2	
78	India	A80	12/13/16	15:42	14	M	1	2	

Figure 10. Filtered Data Table.

At the top center of the dashboard, a dynamic table displays individual-level data for the selected subset. Each row corresponds to a unique respondent and includes demographic variables (country, participant ID, age, and gender), the respondent's self-reported ladder score, and four predictor variables: house quality, food security, access to money, and access to material goods. These features were used as inputs for the predictive models.

3. Model Comparison



Model Comparison

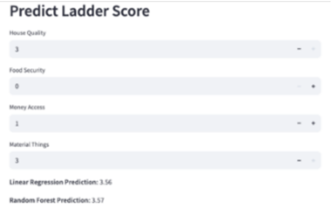
Linear Regression - MAE: 2.27, R^2: -0.01

Random Forest - MAE: 2.43, R^2: -0.19

Figure 11. Model Comparison.

Beneath the data table, two regression models—Linear Regression and Random Forest—are trained to predict children's ladder scores based on reported access to housing, food, money, and material possessions. For each model, performance metrics such as Mean Absolute Error (MAE) and R² are displayed, offering users insight into model accuracy. This side-by-side comparison enables users to evaluate how well different approaches capture variations in subjective social status.

4. Ladder Score Prediction Interface



Predict Ladder Score

House Quality: 3

Food Security: 0

Money Access: 1

Material Things: 3

Linear Regression Prediction: 3.56

Random Forest Prediction: 3.57

Figure 12. Predicted Ladder Score.

Below the model results, an interactive prediction interface allows users to manually input scores for housing, food, money, and things (each ranging from 0 to 3). Based on these inputs, the dashboard displays predicted ladder scores from both regression models. This feature enables users to simulate outcomes and observe how changes in perceived resource access may affect self-reported social status.

5. Logistic Regression Insights

A dedicated section summarizes findings from a binomial logistic regression model. This includes textual interpretation and a color-coded interaction plot showing how the relationship between perceived money access and social status differs by country as shown with Figure 7. For example, while most regions show a negative association between low money access and high status, Argentina displays a reversed trend—underscoring cultural nuance in how children interpret economic standing.

6. Geographic Visualizations

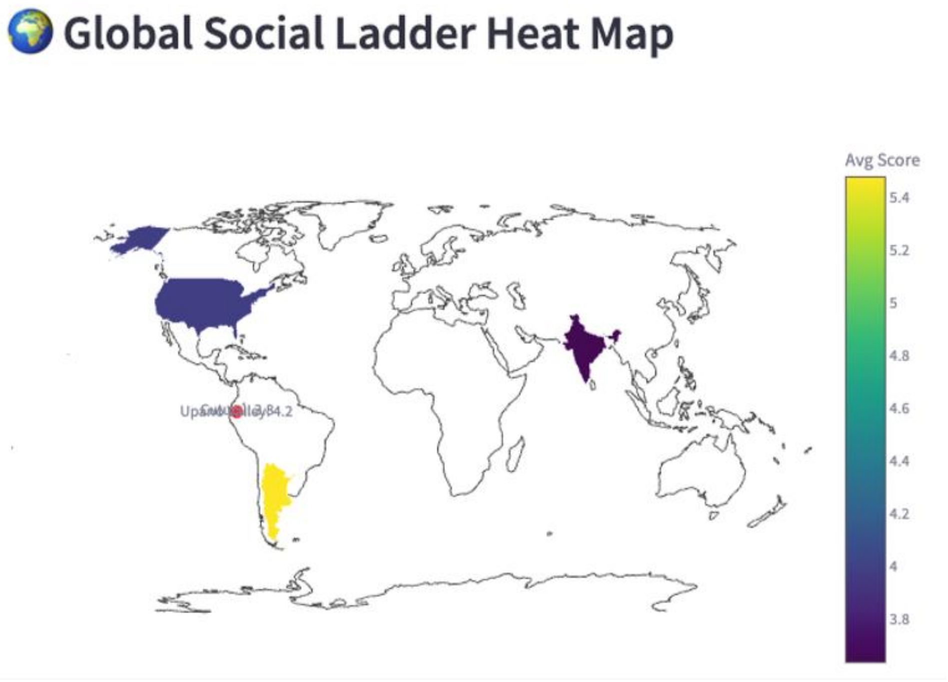
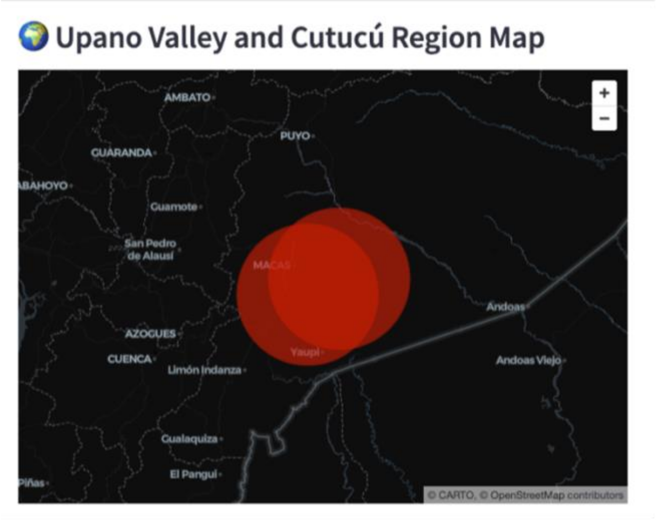


Figure 13. Global Social Ladder Heat Map.



🌍 Upano Valley and Cutucú Region Map

Figure 14. Upano Valley and Cutucu Region Map

Toward the bottom, two visualizations map regional variation in average ladder scores. A choropleth map displays global averages by country, while a Pydeck map highlights specific regions in Ecuador (Upano Valley and Cutucú) with color-coded markers. Together, these tools provide spatial context for interpreting subjective social status across diverse geographic and cultural settings.

Conclusion

Our cluster analysis did not reveal the subgroups that we had hypothesized. The groupings that were discovered were somewhat intuitive: older and younger children with higher or lower subjective social status. Our failure to uncover any more intriguing clusters may be due to the relatively small size of our dataset.

Our logistic regression uncovered an interesting relationship between country and comparative money score in predicting social status. All regions showed that as comparative money score increases, probability of high social status decreases, except for Argentina. This could possibly be due to differences in the cultural importance of money in determining social status as opposed to other metrics.

For Random Forest, our results show that while the model is a powerful tool, its effectiveness depends heavily on the quality of input features and proper tuning. By using age, sex, country, and our engineered material_index as predictors, we were only able to explain about 10% of the variance in stair-climbing ability, with a Mean Squared Error of 6.92 on the test set. These results suggest that the variables we used do not strongly predict the outcome, and the model may be underfitting the data.

Works Cited

Amir D, Valeggia C, Srinivasan M, Sugiyama LS, Dunham Y (2019) Measuring subjective social status in children of diverse societies. PLOS ONE 14(12): e0226550. https://doi.org/10.1371/journal.pone.0226550

DataCamp. (n.d.). *Python tutorial: Streamlit*. DataCamp. Retrieved April 27, 2025, from https://www.datacamp.com/tutorial/streamlit

Ditmars, H. (2024, January 19). Why the ancient Amazonian cities recently discovered in Ecuador are so significant. *The Art Newspaper*. https://www.theartnewspaper.com/2024/01/19/ecuador-amazonian-settlements-lidarupano-valley(The Art Newspaper)

Open Case Studies. (2020, January 27). *Why the struggles of the Shuar Indigenous People in Ecuador to conserve their culture are key to local conservation*. The University of British Columbia. https://cases.open.ubc.ca/why-the-struggles-of-the-shuar-indigenous-people-in-ecuador-to-conserve-their-culture-are-key-to-local-conservation/

Streamlit. (n.d.). *Create an app*. Streamlit Documentation. Retrieved April 27, 2025, from https://docs.streamlit.io/get-started/tutorials/create-an-app

Streamlit. (n.d.). *Streamlit [YouTube channel]*. YouTube. Retrieved April 27, 2025, from https://www.youtube.com/@streamlitofficial

AI Usage

- We used ChatGPT to help generate ideas for creating a heatmap dashboard. Our initial prompt was: *"How can I make a heatmap to showcase my project for a beginner in Python?"* Based on the response, we were introduced to Streamlit as a beginner-friendly tool for dashboard creation. We then consulted the official Streamlit documentation, their YouTube channel and Datacamp documentation to guide our development process. After building the dashboard and reviewing the output, we noticed that specific regions—namely, the Upano Valley and Cross Cutucú—were not highlighted correctly in the heatmap. To troubleshoot this issue, we returned to ChatGPT for assistance in adjusting the heatmap's display to better emphasize these areas.