Life Expectancy Predictor: Modeling Health Outcomes with Structural Indicators

Introduction

Life expectancy is one of the most comprehensive indicators of a population's overall well-being. While traditional analyses often focus on GDP per capita or healthcare spending, these indicators alone rarely capture the underlying structures that shape public health outcomes. This project builds a predictive model using global life expectancy data from the World Health Organization (WHO) to identify deeper structural drivers—those that influence not just how much is available, but how effectively resources are converted into population health. The central task was to create a regression model capable of estimating life expectancy using a carefully selected set of variables. Following detailed exploratory data analysis and extensive model comparison, the strongest performance came from a K-Nearest Neighbors (KNN) regression model using two key predictors: "Income Composition of Resources" and a newly engineered composite metric, "Score_Edu_HIV," calculated by dividing average schooling years by child mortality due to HIV/AIDS. This model achieved a test R^2 score of 0.847 and offered both strong predictive power and interpretability.

To extend the model's usability, I deployed it as an interactive web tool using Streamlit. This interface allows users to input predictor values and instantly receive estimated life expectancy outputs. More importantly, it facilitates real-time feedback, enabling users—whether researchers, policymakers, or general audiences—to assess the accuracy of the model in context and suggest improvements for future iterations.

Data Description

The dataset originates from the WHO Global Health Observatory and includes annual data from 2000 to 2015 for over 190 countries. Variables span health, demographic, and economic domains, including life expectancy, education, mortality rates, immunization, and national income indicators. To ensure reliability and completeness, the analysis focused on the 2014 dataset—the most data-rich year available.

After filtering out rows with missing values in core indicators—life expectancy, schooling, HIV/AIDS mortality in children under five, and income composition—I engineered a new variable: Score_Edu_HIV. This feature captures the balance between a nation's educational reach and its success in mitigating a key health burden. The resulting dataset provided a multidimensional lens on public health structures.

How Features Were Selected

The exploratory phase began with a foundational question: if GDP per capita does not adequately explain life expectancy, then what does? A scatterplot comparing GDP per capita with life expectancy revealed only a weak linear relationship (Pit. 1), suggesting that economic size alone does not guarantee better health outcomes. However, when I compared life expectancy by country classification—developed versus developing—the difference was stark. Developed nations consistently clustered at higher life expectancy levels (Pit. 2), pointing to structural advantages that warranted deeper investigation.

I first tested whether healthcare spending could account for this disparity. The assumption was intuitive: countries that invest more in health should achieve better outcomes.But the data challenged this view. Some nations spent relatively little yet achieved long life expectancies, while others with high expenditures underperformed. The scatterplot made it clear that spending did not equate to system efficiency (Pit. 3).

Shifting focus, I explored education as a structural foundation for well-being. The logic was that higher levels of schooling should improve health literacy, enable better medical decision-making, and promote preventive behaviors. Indeed, a strong, positive correlation emerged between years of schooling and life expectancy—particularly in developing countries (Pit. 4). Education, it seemed, played a vital role in public health outcomes.

This insight led to a second question: does education contribute to life expectancy alone, or are there any other factors? So then, I tested its relationship with income composition—a normalized index that captures access to infrastructure, services, and economic participation. As expected, countries with more schooling tended to have higher income composition scores (Pit. 5), suggesting that income composition could be another factor to contribute to the life expectancy prediction.

Having established that link, I examined whether income composition directly predicts life expectancy. Since this metric reflects a society's ability to provide opportunity and equitable access, it should, in theory, support healthier lives. The results confirmed this expectation. Income composition strongly correlated with life expectancy, especially in developing nations, where access to services and stability can vary widely (Pit. 6).

This, again, raised a follow-up question: beyond education and income composition, could there be another variable that reflects structural well-being? I turned to thinness among youth as a proxy for nutritional status and overall living conditions. If education and income improve access to food and healthcare, then countries with higher levels of both should logically exhibit lower thinness rates. To test this, I analyzed the relationship between schooling, income composition, and thinness. The results confirmed the hypothesis: countries with stronger education systems and higher income composition consistently had lower rates of thinness (Pit. 7; Pit.8), indicating better baseline health and nutrition.

Having validated thinness as an outcome of structural access, I then examined its direct relationship with life expectancy. Since thinness captures not just calorie intake but also the cumulative effects of poverty, food insecurity, and chronic disease, it should function as a strong negative predictor of life span. Indeed, scatterplots showed a clear inverse correlation between thinness and life expectancy (Pit. 9).

Finally, I transitioned from analyzing individual predictors to constructing a composite feature. Recognizing that structural factors interact rather than act in isolation, I created a new variable: Score_Edu_HIV, defined as Schooling divided by HIV/AIDS mortality in children under five. This ratio captures how well a society converts educational investment into disease mitigation. Countries with both high education and low mortality scored well, while those with persistent disease despite schooling scored poorly. Among all features tested, this metric best captured the interaction between social investment and public health outcomes (Pit. 10).

Models and Methods

To build the predictive model, I tested four supervised learning regressors: Linear Regression, Decision Tree Model, and K-Nearest Neighbors (KNN). These models were trained on different combinations of four predictors: Schooling, Income Composition, Thinness, and the Score_Edu_HIV metric. Data preprocessing included standard scaling, and all models were trained on an 80/20 split between training and test sets. I used both R^2 and Mean Squared Error (MSE) to evaluate predictive performance. Each model was tested across different subsets of features to determine the optimal combination and algorithm. Among all tests, the combination of Income Composition and Score_Edu_HIV consistently yielded the strongest results with KNN model. Since KNN emerged as the top-performing model, I optimized its main hyperparameter— k—using GridSearch. The best value was k = 4, delivering a cross-validated R^2 of 0.832 and a test-set R^2 of 0.847. While linear and tree-based models also performed reasonably well, they lacked the consistency and predictive strength of KNN in this context. The neural network model underperformed, likely due to the limited dataset size and absence of deep feature complexity.

Results and Interpretation

The final KNN model with k = 4 and two features (Income Composition and Score_Edu_HIV) proved both accurate and interpretable. Income Composition captures how national wealth translates into opportunity and resource access, while Score_Edu_HIV reveals how effectively education reduces health risks—particularly for vulnerable groups like children. Together, these features produce a model that accurately distinguishes between countries with high structural efficiency and those with gaps in their public health pipeline. Nations scoring high in both indicators consistently achieved longer life expectancy. This model also demonstrates that complex health outcomes don't always require complex models or large numbers of features. In fact, adding more predictors beyond these two did not enhance performance and often reduced clarity. The streamlined nature of this model increases its usability in policymaking and strategic planning.

Beyond the numbers, Score_Edu_HIV offers conceptual power. It acts like a structural efficiency metric: it tells us whether a country's educational investments are translating into measurable reductions in preventable child mortality. High education with high HIV/AIDS mortality suggests systemic disconnect. This metric helps diagnose such breakdowns—informing where public health systems are failing to engage effectively.

Significance

This model challenges the conventional assumption that health outcomes are driven primarily by wealth or spending. Instead, it shows that structure matters. Countries that educate their populations and connect that education to effective health interventions achieve better health—even without the highest GDP.

From a technical perspective, the model demonstrates the value of parsimonious feature selection and thoughtful feature engineering. Score_Edu_HIV was not available in the original dataset—it was created based on theory and context. Yet, it became the most powerful predictor. This underscores how domain insight and creativity can outperform brute-force modeling.

Practically, these findings offer actionable guidance. Public health agencies can use composite metrics like Score_Edu_HIV to assess where structural investments—especially education—fail to reduce disease. In lower-income contexts, this might reflect underfunded school systems or a lack of integration with health outreach. In wealthier settings, it may reveal failure to translate knowledge into behavior change or access. Therefore, this kind of nuance is impossible with metrics like GDP alone. Composite indicators help reveal what's working beneath the surface—and where systems are silently failing.

Looking Forward

While the model performs strongly, future work should expand its temporal and contextual scope. This project only used data from a single year, 2014. In 2025, such a snapshot risks obsolescence. The past decade brought profound changes—COVID-19, economic shocks, and climate disruptions—that likely reshaped health landscapes.

Revisiting this model using longitudinal data (2000–2015) could uncover which countries improved and why. For example, did rising schooling precede falling mortality in certain regions? Did some countries improve health despite stagnant income? Answering such questions would deepen the model's policy value.

Additionally, the success of Score_Edu_HIV suggests a broader strategy: design new composite features grounded in theory. Could we create similar metrics using immunization rates, malnutrition indicators, or even environmental stressors? A richer feature library would expand the model's ability to generalize across health domains.

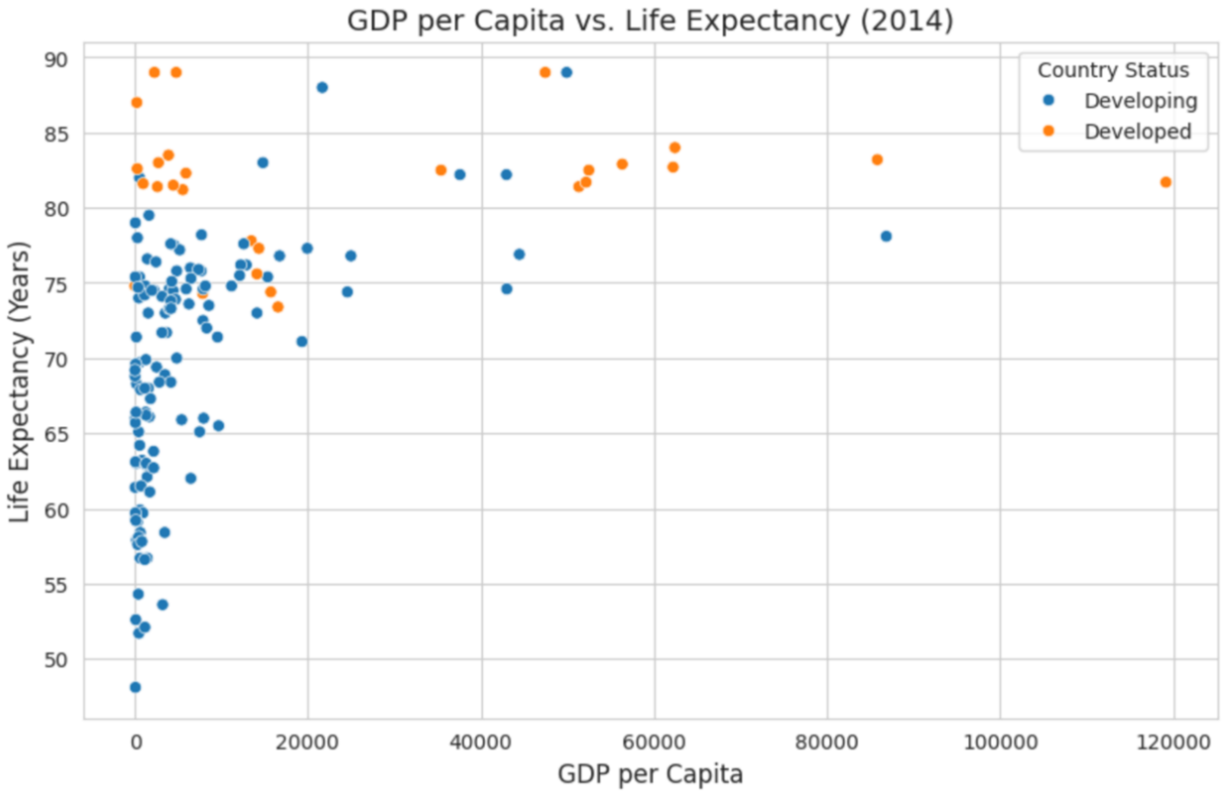
Conclusion

This project began with a deceptively simple question: what predicts life expectancy? Through systematic analysis, thoughtful feature engineering, and rigorous modeling, it produced a surprisingly powerful answer: structural efficiency. Life expectancy improves not just with spending, but with societies that educate their citizens and connect that education to real-world health outcomes.

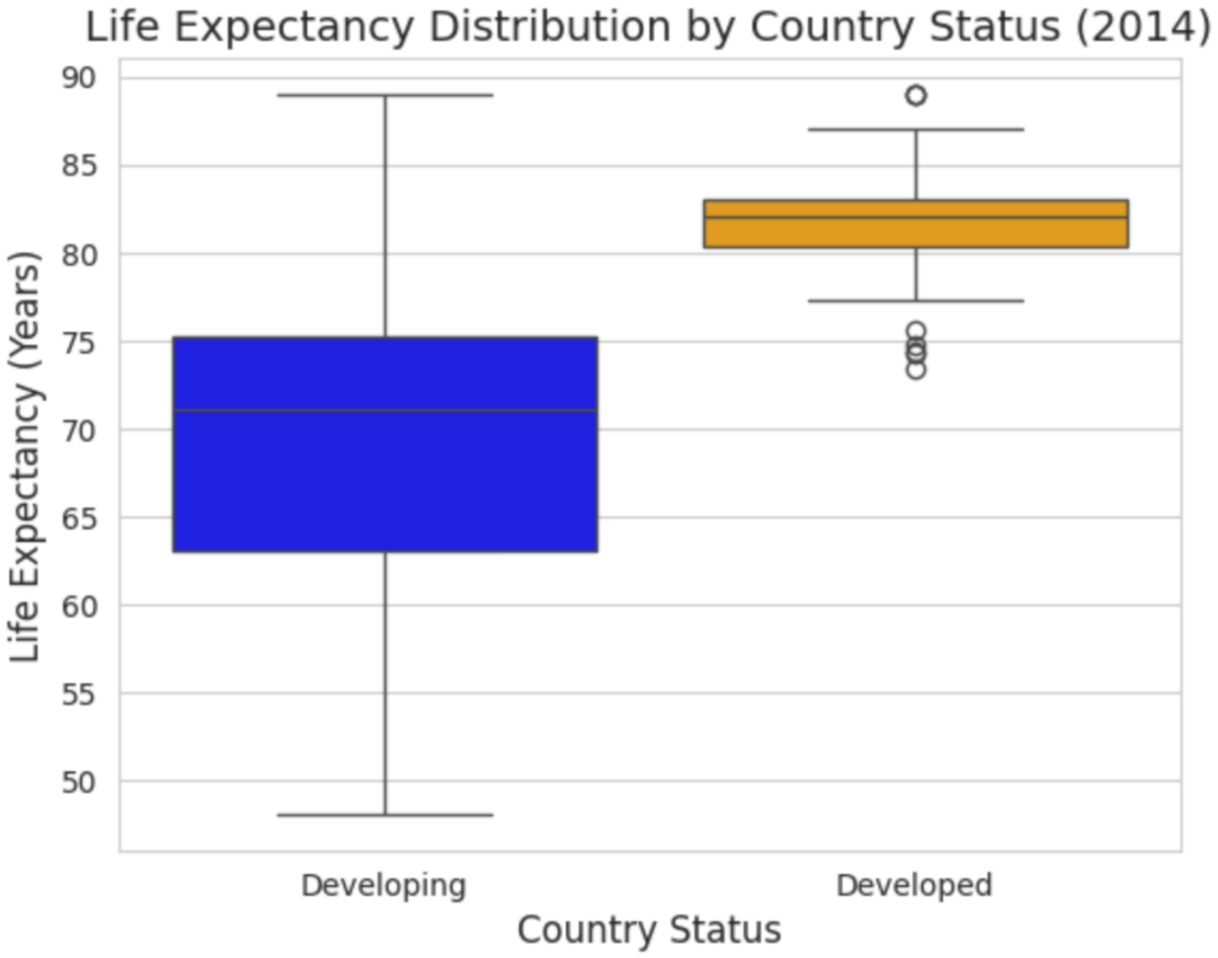
The accompanying Streamlit app provides a platform for engagement. Users can adjust input values, receive predictions, and reflect on the real-world implications. More importantly, they can offer feedback—helping guide future iterations of the model based on new knowledge, user experience, and regional nuance.

Ultimately, this project illustrates how data can reveal truths behind national health performance—and how combining numbers with context can lead to more equitable futures.

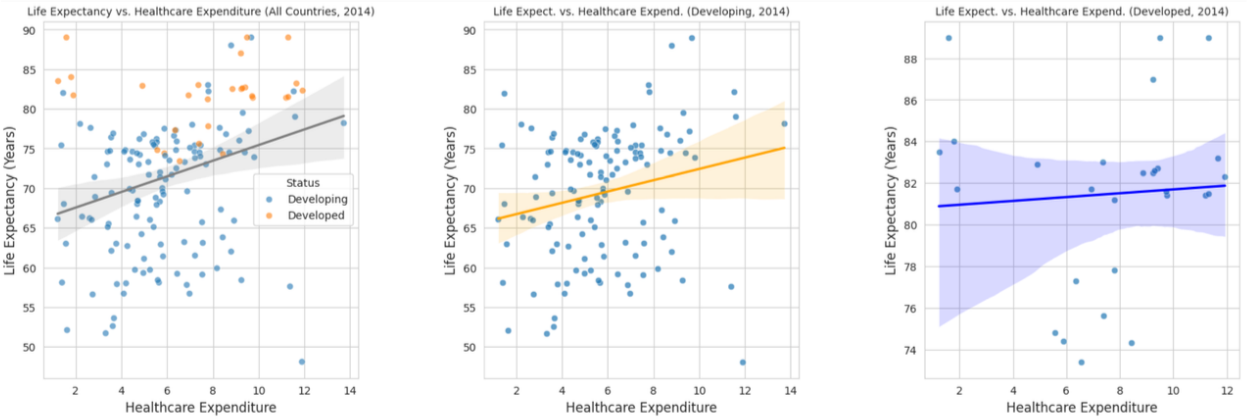
Attachment



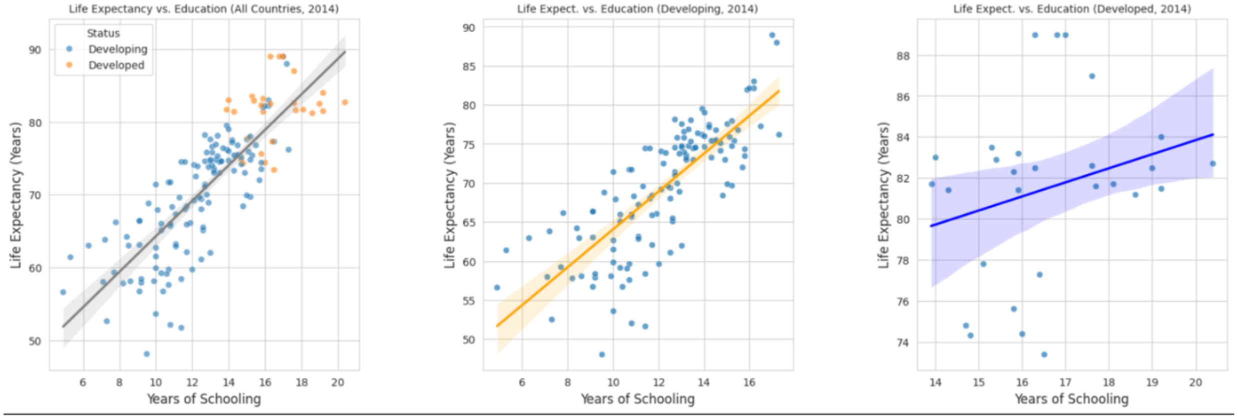
(Pit.1 Scatter Plot GDP per Capita vs. Life Expectancy)



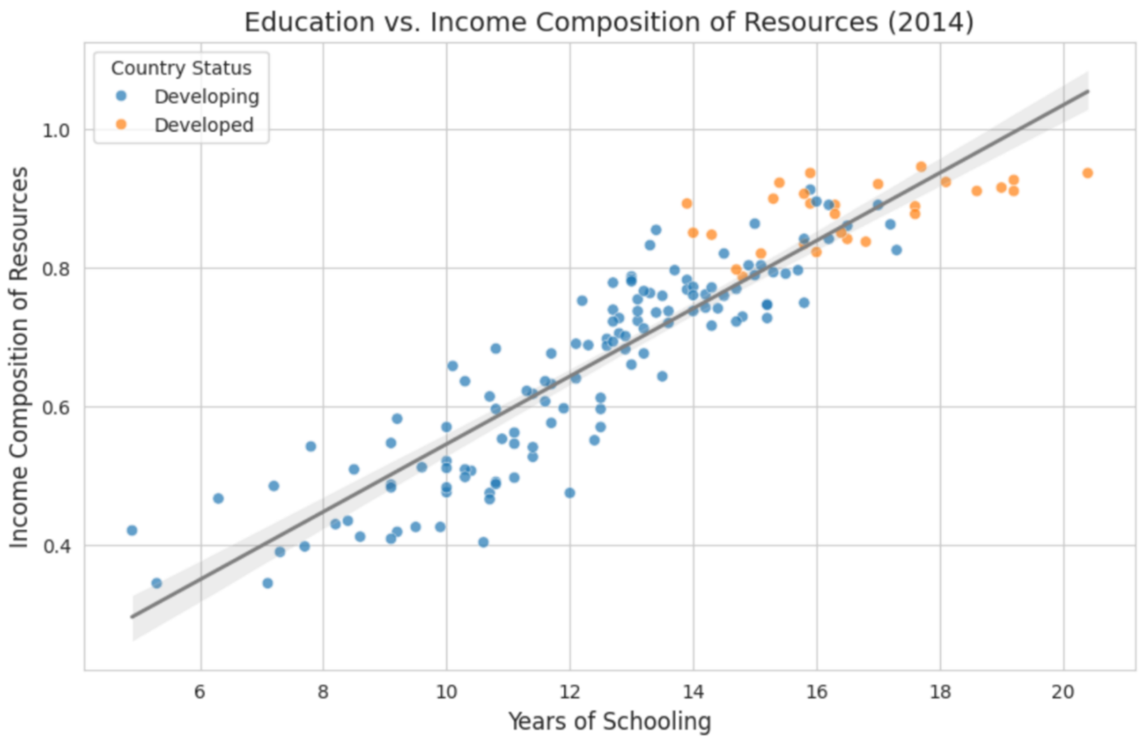
(Pit.2 Box Plot of Life Expectancy by Country Status)



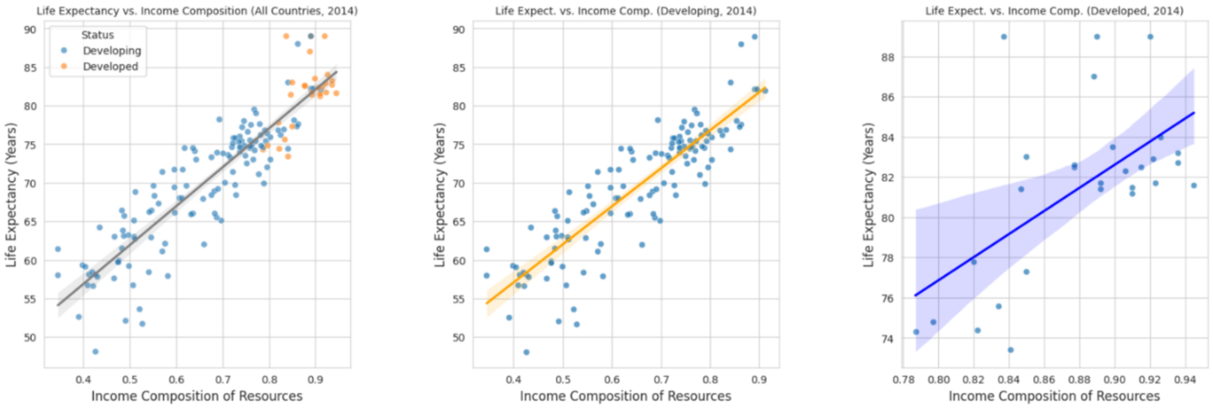
(Pit.3 Scatter Plot Healthcare Expenditure vs. Life Expectancy)



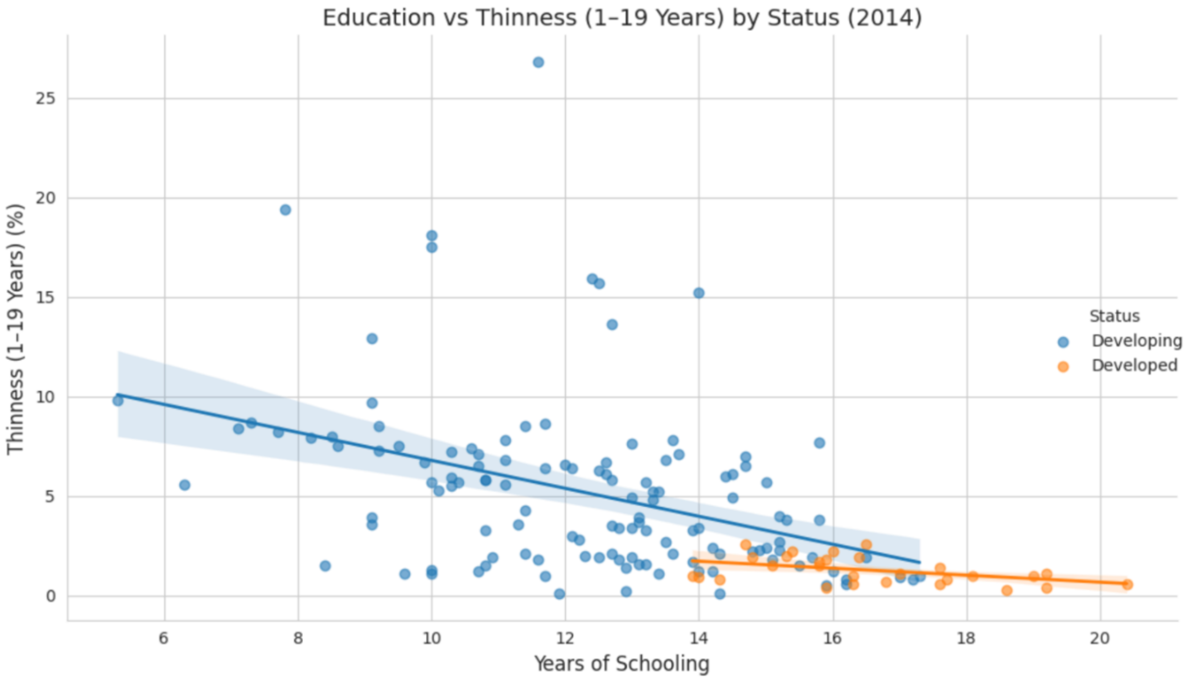
(Pit.4 Scatter Plot Education vs. Life Expectancy)



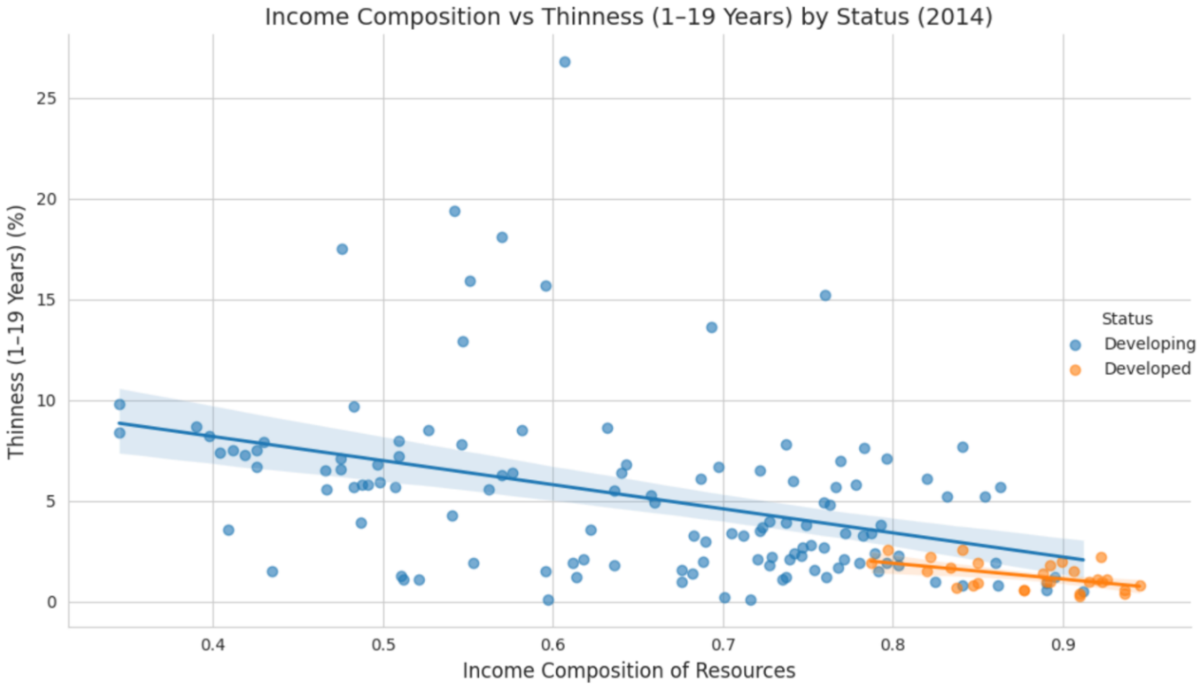
(Pit.5 Scatter Plot Education vs. Income Composition)



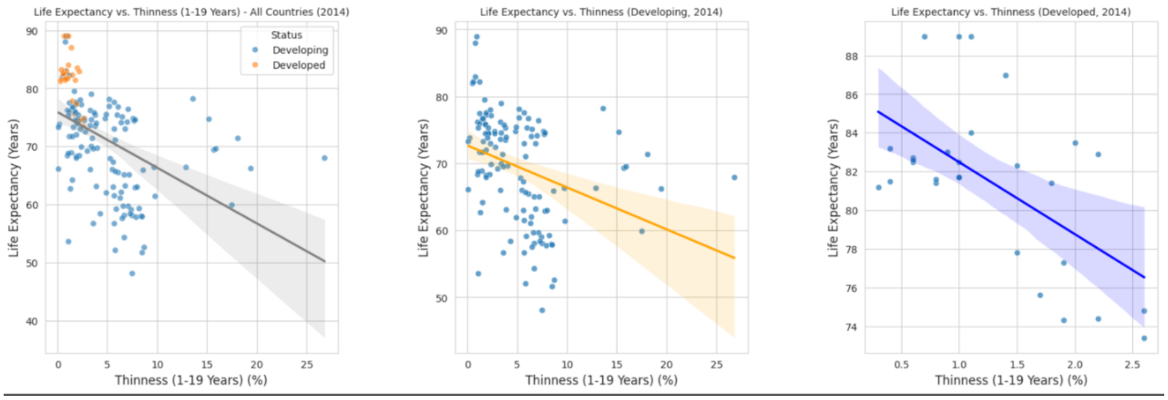
(Pit.6 Scatter Plot Income Composition vs. Life Expectancy)



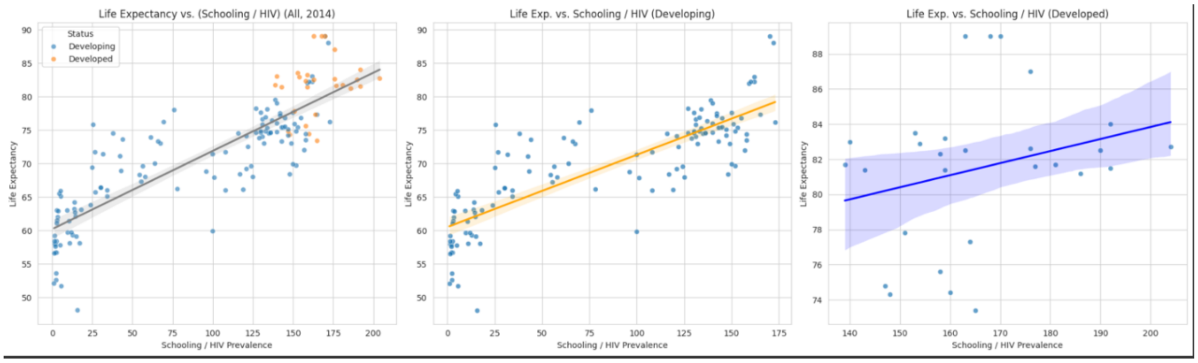
(Pit.7 Scatter Education vs. Thinness)



(Pit.8 Scatter Plot Income Composition vs. Thinness)



(Pit.9 Scatter Plot Thinness vs. Life Expectancy)



(Pit.10 Scatter Plot Schooling / HIV vs. Life Expectancy)