Jimmy Chen
Data Analytics Final Report
CSCI 4600 (4000 Level)
Dr. Ahmed Eleish
chenj62@rpi.edu

Analyzing Global Development:
Linking Economic Wealth with Health and Education Outcomes

Abstract & Introduction

Economic wealth can be viewed from many different perspectives. For politicians, it often means national GDP growth, but for households, it can mean income and assets. Gary S. Becker notes that “to most people, capital means a bank account, a hundred shares of IBM stock, assembly lines, or steel plants in the Chicago area” (Becker 1). In Becker’s study, he mentions that the tangible forms of capital like money isn’t the only type of capital. There’s other forms like schooling and medical care expenditures. While traditional economic assessments often have a huge impact on GDP, alternative indicators such as health expenditue, literacy rates, and poverty levels offer deeper insights into true economic well-being of populations. Becker mentions that investments in this area is often referred to human capital, and the reason being is that, “people cannot be separated from their knowledge, skill, health, or values in the way they can be separated from their financial and physical assests” (Becker 1). He emphasizes the importance of investing in human capital, as it’s ultimately the biggest contributing factor to a country’s development.

Building off of Becker’s study, this project will explore the relationships between the GDP per capita and several socioeconomic indicators to understand how they affect each other and the type of predictions we can make from them. Our goal is to understand the genuine influence each indicator has on a country's wealth and determine whether these factors alone can help predict future economic conditions. The motivation behind this project comes from interest to test common beliefs about education's importance in driving economic growth, as well as to validate the idea that good health significantly contributes to development, since healthy populations are essential for sustainable progress. Understanding these relationships could drive more effective policy decisions and also provide meaningful insights for strategies in poverty alleviation, improvement in public health, and overall a stronger economy.

Data Description & Preliminary Analysis

The dataset used in this project is sourced from the World Development Indicators (WDI) from the World Bank DataBank. Specifically within the DataBank, we use specific indicators, including the GDP per capita (USD), poverty headcount ratio, life expectancy, infant mortality, health expenditure per capita, adult literacy rate, and primary school enrolment ratio. The criteria for selecting these indicators were their relevance to measuring aspects of human capital thoroughly. For example, indicators such as life expectancy, infant mortality rates, and health expenditure per capita, collectively provide a thorough representation of a nation's overall health status. World Bank release a new R package in March 22, 2025, so data retrieval was simple and clean as we only had to specify the indicator abbreviations, the year range, and the countries. All of the data we pulled were numeric values, which made it super easy to work with. Each row represented a new year, and each country had multiple years ranging from 2000-2023. Originally the plan was to use data starting from the 1960s, but based on further manual analysis, most of the data collection for some indicators only started in 2000s. The countries were mostly individualized, but sometimes, when there wasn't enough data, they were grouped into geographical sections like "East Asia & Pacific".

Preliminary analysis of the data involved creating histograms and boxplots to assist with understanding the data before preprocessing. The R library "skimr" (Figure 1) was also utilized to quickly give a rundown of the number of missing values for each indicator, as well as the standard deviation, mean, and other statistics. From this data, it can be inferred that there are large gaps in GDP between countries due to it's large standard deviation. Two indicators that really stood out was that life expectancy across countries was generally similar, but the health expenditure for the countries show a lot more disparity. This could show that the two indicators don't really have an affect on each other as we originally thought.

```
── Data Summary ────────────────────────
                           Values
Name                       df_analysis %>% select(ye …
Number of rows             6384
Number of columns          9
_____________________
Column type frequency:
  character                1
  numeric                  8
_____________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 iso3c               120         0.981   3   3     0      261          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────
  skim_variable     n_missing complete_rate     mean       sd      p0     p25     p50     p75     p100 hist
1 year                      0         1      2012.       6.92 2000    2006.   2012.   2017.    2023  ▇▇▇▇▇
2 ny_gdp_pcap_cd          205         0.968 14639.   23088.    110.   1536.   5003.  18357.  256581.  ▇▁▁▁▁
3 si_pov_dday            4462         0.301     9.18    15.6     0      0.2     1.4    11.1     91.5 ▇▁▁▁▁
4 sp_dyn_le00_in           24         0.996    70.5      8.58   14.7    65.2    72.0    76.9     86.4 ▁▁▃▅▇
5 sp_dyn_imrt_in          528         0.917    28.0     24.9     1.4     8      19.3    42.9     235. ▇▂▁▁▁
6 sh_xpd_chex_pc_cd       923         0.855   932.    1673.      4.18   63.7   248.    806.    12434. ▇▁▁▁▁
7 se_adt_litr_zs         4556         0.286    79.6     17.2    14      66.0    84.0    94.3     100  ▁▁▃▅▇
8 se_prm_enrr            1248         0.805   102.      12.8     8.45   97.4   102.    107.     257.  ▇▇▁▁▁

indicator keys:
"ny_gdp_pcap_cd",     # GDP per capita (current US$)
"si_pov_dday",        # Poverty headcount ratio $1.90/day
"sp_dyn_le00_in",     # Life expectancy at birth (years)
"sp_dyn_imrt_in",     # Infant mortality (per 1,000 live births)
"sh_xpd_chex_pc_cd",  # Current health expenditure per capita
"se_adt_litr_zs",     # Adult literacy rate (% ages 15+)
"se_prm_enrr"         # Primary school enrolment ratio (% gross)
```

Figure 1: Skimr summary of data before pre-processing and cleaning

The histograms (Figure 2) showed a extreme right skewed distribution for GDP per capita, with most of the countries together at lower income levels, and a longer tail extending towards very high incomes. There's a similar skew to this within the health expenditure per capita, where a few countries spend way more on health than the majority. On the contrary, percentage-based indicators like the adult literacy rate and primary school enrollment cluster toward their maximum values. Life expectancy is roughly bell-shaped, centered around 70 years, but with a tail of lower values reflecting countries with high mortality. Infant mortality and poverty headcount are strongly right-skewed as most countries have low infant mortality and low extreme poverty. The boxplots (Figure 3) further support this data, with GDP per capita and health expenditures having the most outliers and broad ranges, and literacy and life expectancy with shorter ranges. These plots highlight the diversity in development outcomes, as many distributions deviate from normality, showing that it might be harder to predict exact outcomes than originally planned.

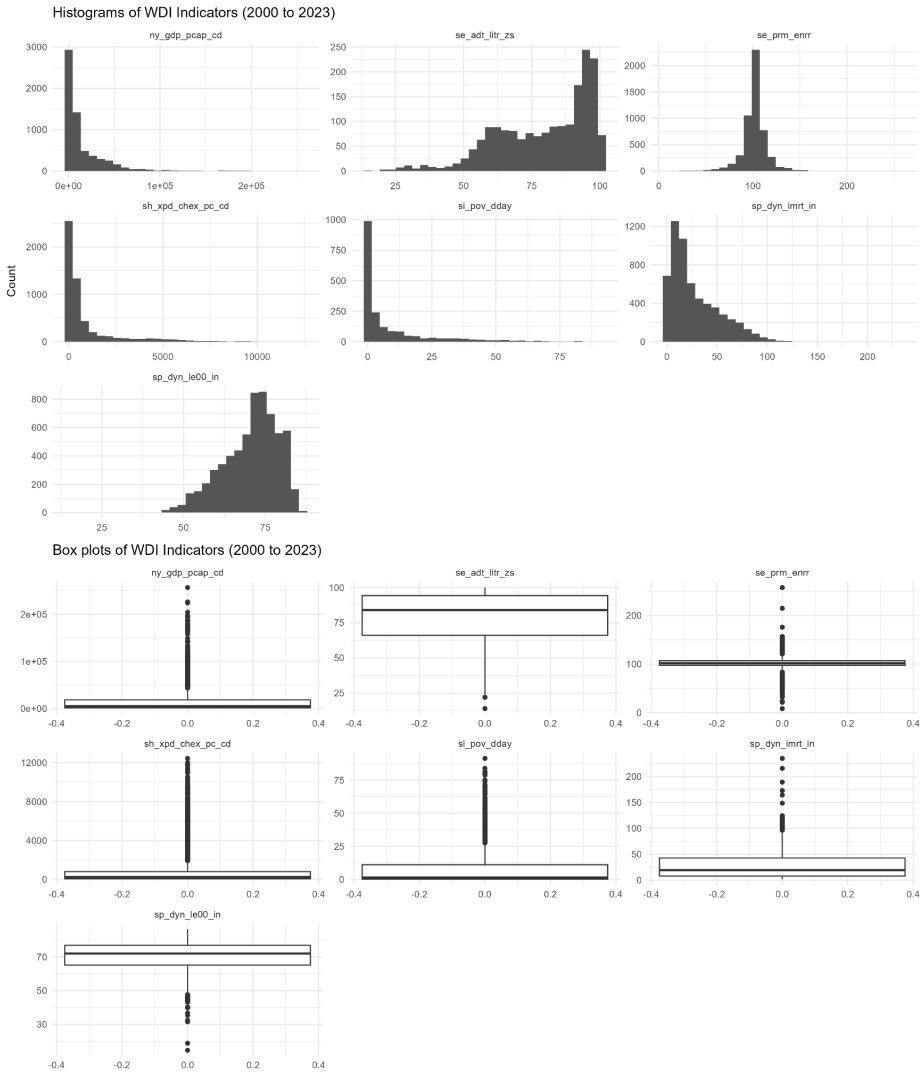

Figure 2 & 3: Histograms & Boxplots of indicators before pre-processing and cleaning

Exploratory Analysis

The WDI dataset with our custom indicators of GDP per capita (USD), poverty headcount ratio, life expectancy, infant mortality, health expenditure per capita, adult literacy rate, and primary school enrolment ratio, went through a large amount of transformation, smoothing and cleaning before analysis. First for easier readability, indicators were renamed from the abbreviations to the actual titles. Then, as mentioned in the preliminary analysis, we also had to shorten the data due to the large amount of NA values. Even after limiting the data to 2000-2023, there was still over 10,000 values missing in total (Figure 1), as not all countries reported every indicator annually.

To address these missing values problem and prevent bias from improper calculations, we used interpolation to fill in missing data like the health expenditure based on its' existing trends, and in some cases the median of neighboring years. This method preserved each country's overall trajectory and avoided artificially inflating or deflating values, which helped make sure that we were using complete data for our models. By the end of cleaning, each variable was then normalized into a Z-score after applying the transformations, which made the data more comparable and removed any extreme outliers. The skewed monetary indicators like GDP per capita and health expenditure per capita were also log-transformed before standardization to reduce right-skewness. Applying the log transformation brings larger values closer to the rest of the data, preventing a the few wealthy countries from excessively influencing the analysis. These preprocessing steps such as handling missing values, log-transforming skewed indicators, and standardizing helped prepare the dataset for proper analysis and model training.

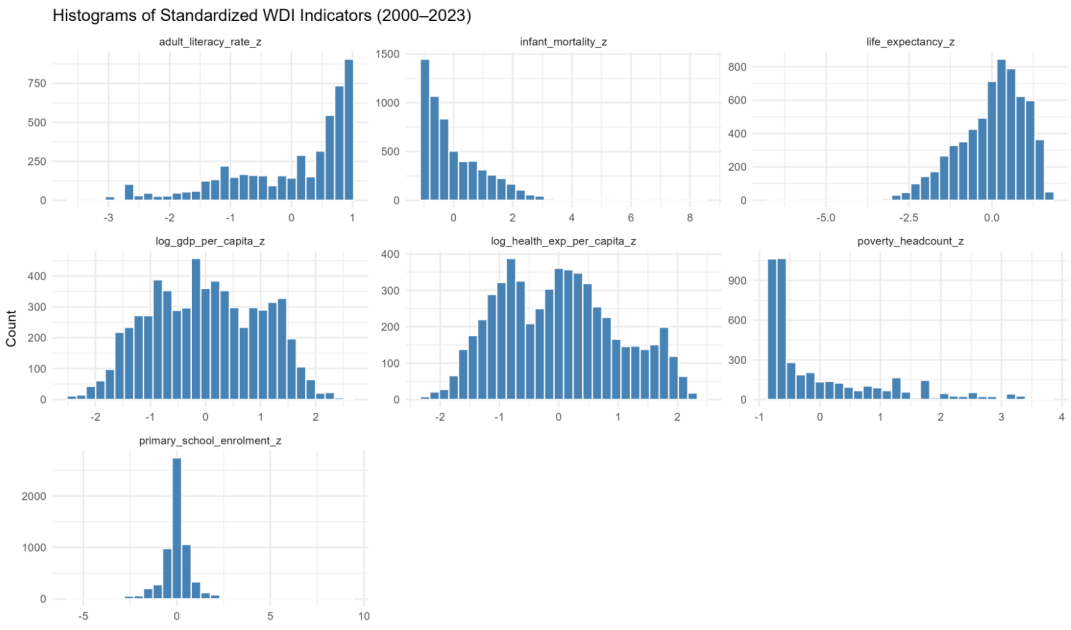

Figure 4: Histograms of indicators after standardization and cleaning

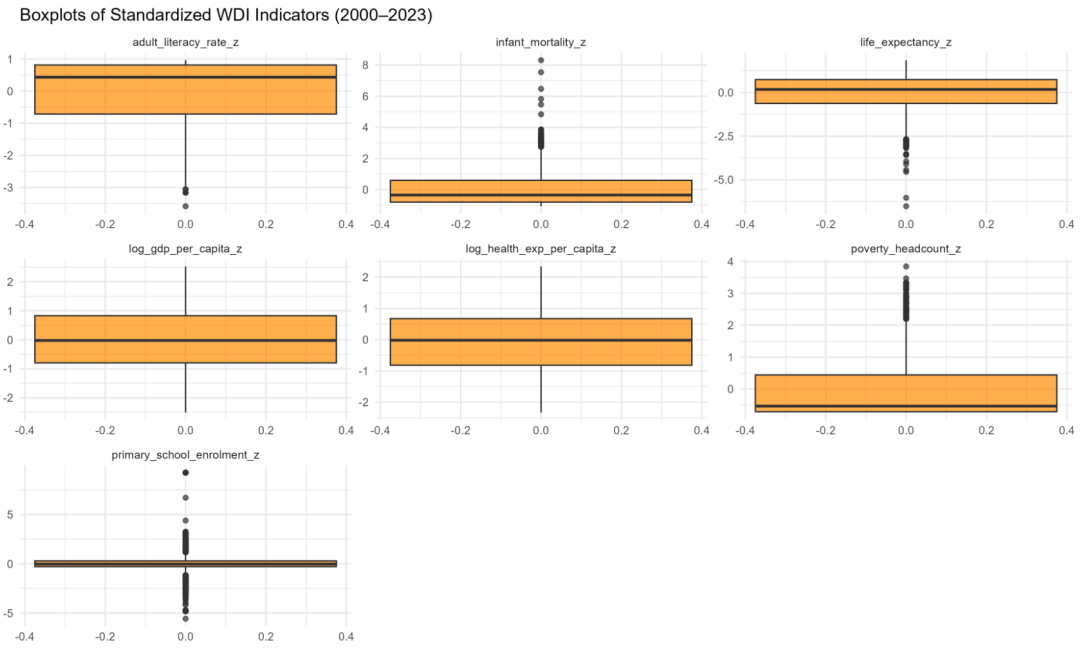

Figure 5: Boxplot of indicators after standardization and cleaning

Figures 4 and 5 above are the histograms and boxplots of the same indicators after all the cleaning and processing was done. The histograms now show a lot more of a symmetric distribution, with most countries around the mean, and fewer extreme deviations/outliers. Poverty and mortality still remains right skews, mainly because they're low percentages, but we can conclude that their tails are shorter than before.

After cleaning it was easier to have a closer and deeper analysis of our dataset and also find the relationships among the variables. Our data shows that the log GDP per capita has a mean of 0 in standardized units, with a standard deviation of 1, which corresponds to an average of $10,000-$20,000 in USD if we backtrack the transformations. Life expectancy was roughly 70 years, infant mortality was roughly 25 per 1000 births, and health expenditure showed the biggest jump in spending among recent years. Poverty headcount is low in most countries, but the mean is pulled up by a couple low-income countries. Some countries have 40–60% of the population in extreme poverty, which translates to poverty_Z scores of +2 or higher. These summary statistics highlights the development gap between countries, as many nations achieve high life expectancy, low infant mortality, low poverty, and nearly full literacy, while others fall far behind on these metrics, which often corresponds to much lower GDP per capita.

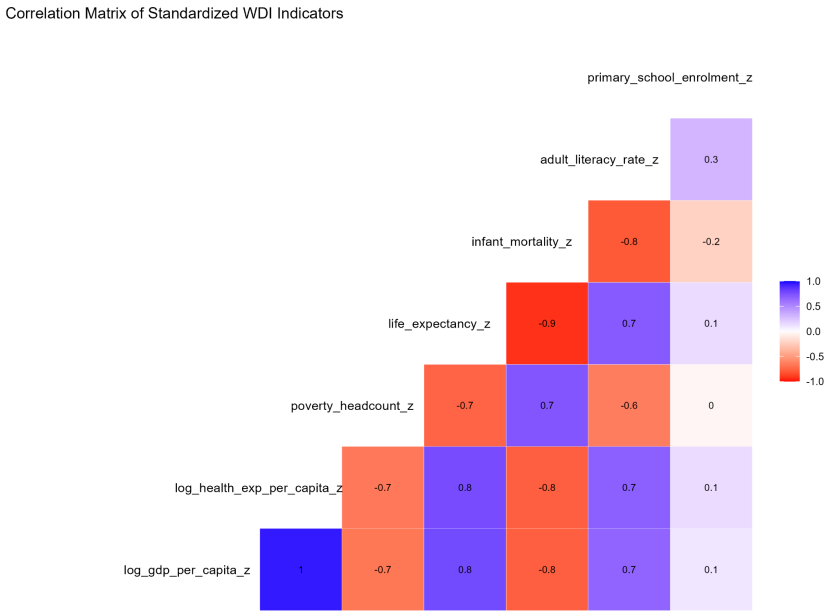

Figure 6: Correlation matrix of each of the WDI indicators

We were able to fully confirm these by creating a correlation matrix (Figure 6), which showed us that Log GDP per capita stands out with strong positive correlations to log health expenditures per capita, and life expectancy, both at r = 0.8. There was also negative correlation between Log GDP per capita and infant mortality and poverty headcount, at r = -0.8 and r = -0.7 respectively. This confirms that richer countries tend to have higher health spending, longer-lived populations, and lower poverty and infant mortality. Primary school enrollment actually had rarely any correlation with the other indicators at r = 0.1 with GDP, but mostly because most countries all already had primary school enrollment regardless of income and health. This could bring some uncertainty to our hypothesis, as adult literacy corresponds closely with GDP at r = 0.7, so we don't know the true effect of education. It could be assumed that not everyone who enrolls in primary school pursues higher education, but it is still an uncertainty. In terms of bias, when we address missing values, we made sure to target most of the NA values, but not forcefully fix all of them. For example, during times like if a country never reported literacy rates, they would be excluded from the correlations involving literacy, because it would introduce bias if we just estimate their literacy rates.

Overall, after cleaning and all the analysis we conducted, it can be concluded that countries that are prosperous tend to also have better health outcomes and generally educated populations and vice versa. These findings sets the stage when conducting our predictive modeling, where we will examine how well we can predict GDP per capita from health and education indicators, and which factors shows up as the most influential.

Model Development and Application of Model(s)

The choice of model for our prediction is very influential of the results we get, so it was important to understand our data before we started modeling. The strong correlations between indicators suggests that a linear model might be able to capture a large portion of the variance in GDP, but at the same time, using non linear relationships could yield more results as we are able to reference indicators in more flexible ways. The models that we ended up developing was a linear regression model and a random forest regression. The linear model would let us have a straightforward interpretation of coefficients, while the random forest will be able to capture non-linear effects and rank the importance of each predictor. Together, these approaches will help us understand not just the extent to which health and education factors can explain differences in GDP, but also which specific factors carry the most weight and how they might interact.

The target variable for both models was the log of GDP per capita. The predictors considered were the full set of indicators mentioned earlier such as health expenditure per capita, life expectancy, infant mortality, adult literacy, primary enrollment, and poverty rate, which were all standardized from before. Both models were trained on the combined data from 2000 - 2023, treating each country-year as an independent observation. While this ignores autocorrelation and country-specific effects, it increases the sample size to around N = 4500 observations which allows the models to utilitze the variability across countries and over time. We accounted for overfitting and avoided generalization by using 5-fold cross-validation. The data were randomly split into five subsets, and the models were trained on 80% and tested on the 20%. This was repeated so every point was in a test fold at least once. This approach provides a well-rounded estimate of how well the models are predicting and performing.

Model tuning was done within the cross-validation loop to avoid overly optimistic bias. For the random forest, we did a grid search over key hyperparameters (the number of variables randomly sampled at each split, mtry, and the minimum node size) to find the combination that minimized prediction error on validation folds. This did end up taking a while to run though, due to the amount of different combinations there were. The linear regression had no hyperparameters beyond the choice of predictors, but we did examine the statistical importance of coefficients and potential multicollinearity. If any predictors were found to be redundant or highly collinear, we made sure to address it. For example, because life expectancy and infant mortality were so strongly correlated, it was expected that one of them might dominate in the linear model, and so we remained cautious in interpreting the coefficients. The random forest model, was way less susceptible to multicollinearity but can still be influenced by very strong predictors overshadowing others.

Diving deeper into the multiple linear regression model, the model used log GDP per capita as the dependent variable and the suite of health, education, and poverty indicators as independent variables. Initially both life expectancy and infant mortality were included with the awareness that their high correlation might cause instability, but the aim was to let the regression determine which of these significantly contribute when controlling for others. The model was fitted to the data using ordinary least squares, and the resulting fit was pretty high in terms of explained variance.

model	rmse	rsq	mae
<chr>	<dbl>	<dbl>	<dbl>
Linear	0.230	0.917	0.171

Figure 7: Linear Regression Model RMSE, RSQ, and MAE

The model achieved an R^2 of about 0.92 on average in cross-validation and 0.93 on the full sample, which means it explains roughly 92% of the variance in log GDP per capita (Figure 7). The cross-validated Root Mean Squared Error (RMSE) was around 0.23 in log GDP units. Since the target is log-scaled, an RMSE of 0.23 corresponds to a prediction error of about 23% in GDP, which is one standard deviation of log GDP. The Mean Absolute Error (MAE) was approximately 0.17 in log units, which is about a 17% error on GDP. It was concluded that these error rates are pretty low given the diversity of countries and years. The high R^2 indicates that a linear combination of these indicators captures most of the variation in economic wealth, strengthening our hypothesis that health and education factors, alongside poverty, are fundamentally tied to economical development.

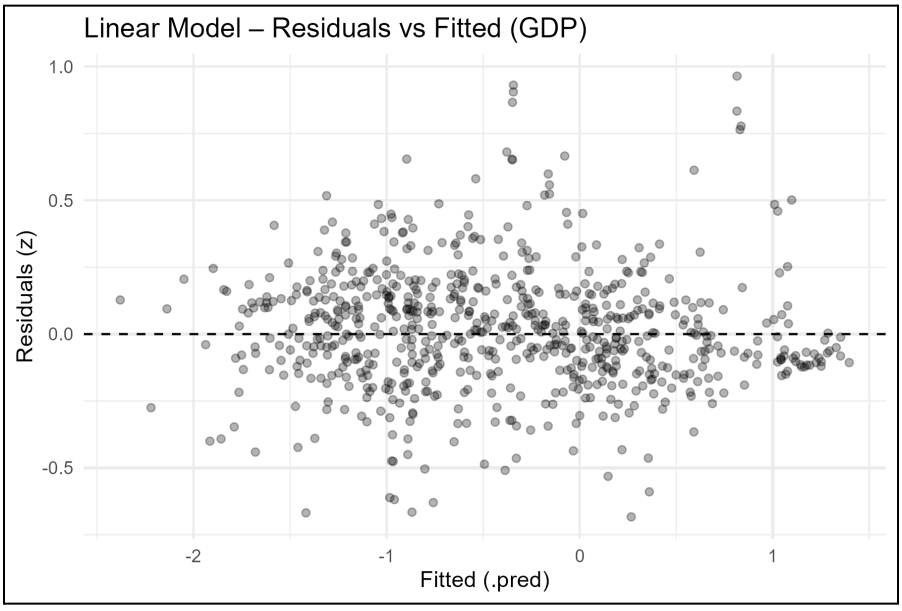

Figure 8: Residuals vs. Fitted Values for the linear regression model predicting log GDP

Figure 8 shows a plot of the residuals versus the fitted values for log GDP per capita. Each point represents a country-year observation, with the x-axis being the model's predicted GDP (fitted value) and the y-axis as the residual (actual - predicted gdp). The dotted horizontal line at 0.0 indicates perfect prediction. The residuals are distributed roughly symmetrically around that line, with no clear trends. There also is no funnel shape or systematic curve, which means that the variance of the residuals are relatively constant. There are a few outliers visible, like all the ones around -0.5 to -1.0 on the left side, which means that the model over-predicted GDP for those observations. This could be a result of some countries having lower GDP than expected, given their health/education indicators. However, these outliers are only a few, in comparison to the dataset size, and overall, the linear model form appears pretty accurate for capturing the bulk of the relationships.

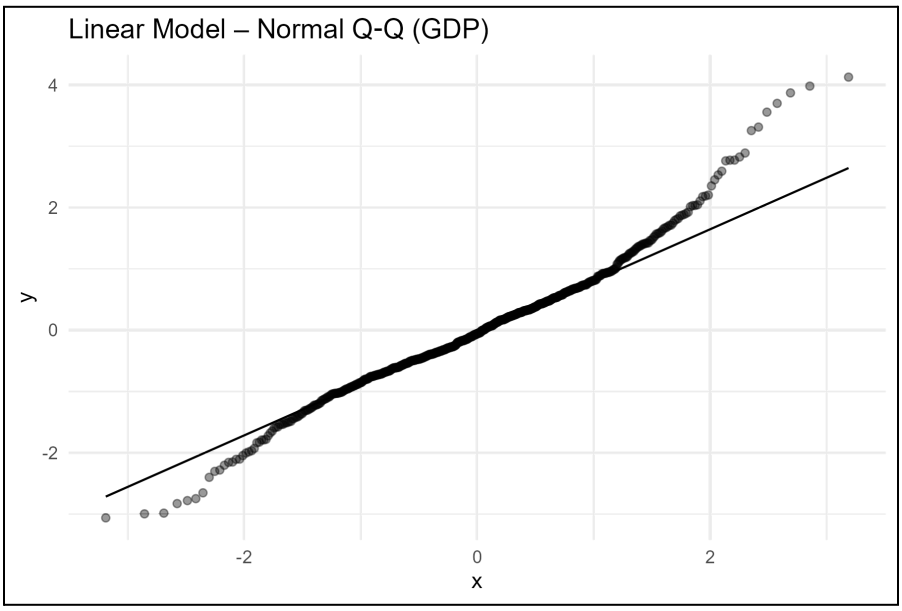

Figure 9: Normal Q–Q plot of the linear model residuals

The normality of the residuals were also analyzed using a Q-Q plot to see if it followed a normal distribution (Figure 9). The residuals lie almost right on top of the reference line, especially within the middle range of -2 to 2 on the standardized residual scale. There was a slight deviation at the extreme right end, with the largest positive residuals, which represents a few country-years where the GDP was under-predicted. The same is said for the extreme left tail, which shows a minor divergence as well. These deviations are not uncommon in cross-country economic data, as they may correspond to other factors like for example, perhaps a oil-rich economy has higher GDP, but doesn't invest much in healthcare, so the predictions on that economy might be under-represented. Overall, we can see that the residuals conform to normality, reinforcing the confidence intervals and p-values for the regression coefficients.

	mtry	min_n	.metric	.estimator	mean	n	std_err	.config
	<int>	<int>	<chr>	<chr>	<dbl>	<int>	<dbl>	<chr>
1	2	3	rmse	standard	0.273	5	0.0176	Preprocessor1_Model01
2	2	3	rsq	standard	0.888	5	0.0166	Preprocessor1_Model01
3	3	3	rmse	standard	0.268	5	0.0178	Preprocessor1_Model02
4	3	3	rsq	standard	0.893	5	0.0159	Preprocessor1_Model02
5	4	3	rmse	standard	0.267	5	0.0182	Preprocessor1_Model03
6	4	3	rsq	standard	0.894	5	0.0159	Preprocessor1_Model03
7	6	3	rmse	standard	0.269	5	0.0185	Preprocessor1_Model04
8	6	3	rsq	standard	0.891	5	0.0167	Preprocessor1_Model04
9	2	8	rmse	standard	0.273	5	0.0172	Preprocessor1_Model05
10	2	8	rsq	standard	0.889	5	0.0161	Preprocessor1_Model05
11	3	8	rmse	standard	0.268	5	0.0174	Preprocessor1_Model06
12	3	8	rsq	standard	0.893	5	0.0158	Preprocessor1_Model06
13	4	8	rmse	standard	0.266	5	0.0178	Preprocessor1_Model07
14	4	8	rsq	standard	0.894	5	0.0158	Preprocessor1_Model07
15	6	8	rmse	standard	0.268	5	0.0184	Preprocessor1_Model08
16	6	8	rsq	standard	0.892	5	0.0166	Preprocessor1_Model08
17	2	14	rmse	standard	0.274	5	0.0167	Preprocessor1_Model09
18	2	14	rsq	standard	0.888	5	0.0161	Preprocessor1_Model09
19	3	14	rmse	standard	0.267	5	0.0169	Preprocessor1_Model10
20	3	14	rsq	standard	0.894	5	0.0153	Preprocessor1_Model10
21	4	14	rmse	standard	0.266	5	0.0178	Preprocessor1_Model11
22	4	14	rsq	standard	0.894	5	0.0157	Preprocessor1_Model11
23	6	14	rmse	standard	0.268	5	0.0181	Preprocessor1_Model12
24	6	14	rsq	standard	0.892	5	0.0164	Preprocessor1_Model12
25	2	20	rmse	standard	0.273	5	0.0163	Preprocessor1_Model13
26	2	20	rsq	standard	0.889	5	0.0159	Preprocessor1_Model13
27	3	20	rmse	standard	0.266	5	0.0165	Preprocessor1_Model14
28	3	20	rsq	standard	0.895	5	0.0154	Preprocessor1_Model14
29	4	20	rmse	standard	0.265	5	0.0175	Preprocessor1_Model15
30	4	20	rsq	standard	0.895	5	0.0154	Preprocessor1_Model15
31	6	20	rmse	standard	0.267	5	0.0180	Preprocessor1_Model16
32	6	20	rsq	standard	0.893	5	0.0162	Preprocessor1_Model16

Figure 10: Full grid of 16 combinations' metrics for Random Forest

The Random Forest model works by building a large number of decision trees on random subsets of data, and then averaging their predictions. This generally leads to higher accuracy and prevents overfitting, which is what we wanted to test along with the linear model. The random forest was trained on the same cross-validation setup as the linear model, and the parameters were turned with a grid search, as mentioned in the beginning summary of this section. More specifically, the key parameters turned were mtry and the minimum node size. We tested values from 2 to 6 for mtry, and found that a middle mtry of around 4 worked the best, as it yielded the lowest cross-validated RMSE of 0.267 (Figure 10). We also tuned the minimum nodes and found an optimal value that balanced bias and variance (a leaf minimum of about 3 data points gave best results). We set the number of trees to 1000, which was sufficient for the model to stabilize.

model	rmse	rsq	mae
<chr>	<dbl>	<dbl>	<dbl>
Linear	0.230	0.917	0.171
Random-Forest	0.155	0.962	0.113

Figure 11: Random-Forest Model Statistics Addition to Figure 7

The performance of the random forest higher than the linear regression, with the RMSE at around 0.96, which means that the model explains about 96% of the variance in log GDP per capita. In correspondence to that, the RMSE dropped to around 0.155 in log units, and MAE to around .11. This means that there could be some non linear patterns in the data that the linear regression wasn't able to capture. One possibility could be diminishing returns, like for example, if there was an increase in life expectancy from 50 to 60, it might have a greater affect on GDP than an increase from 60 to 70.

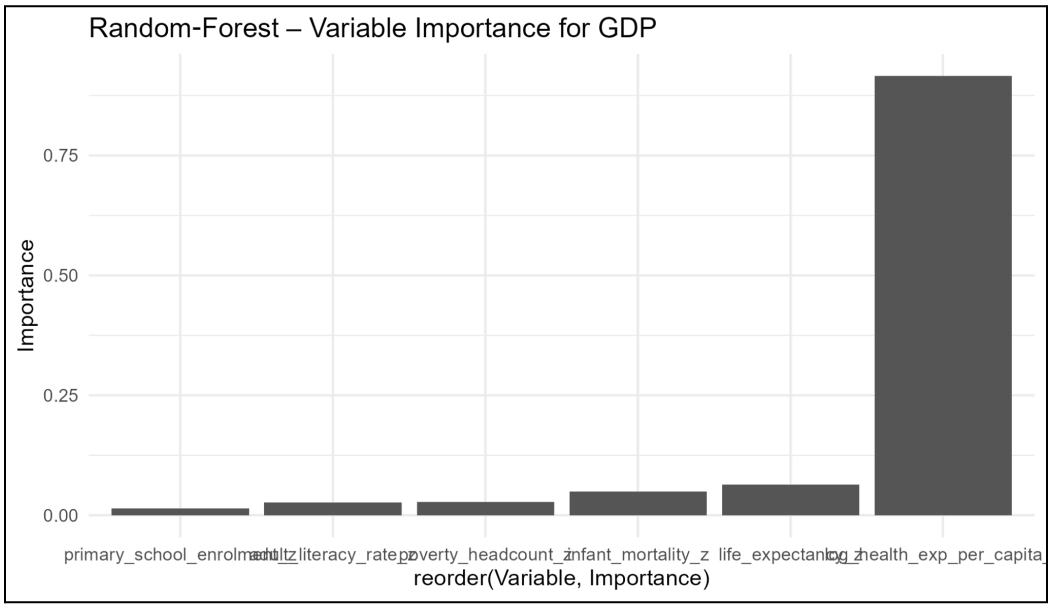

Figure 12: Random Forest Variable importance plot for GDP

Figure 12 shows the importance of each indicator, and paints a picture of which indicator was the most useful for prediction. The greatest one was health expenditure, which was around 10 times more important than all the other indicators. The difference was so large that we could almost just use health expenditure for prediction. One major reason that could be the causing factor for this is that many times, GDP itself can often influence how much countries spend on health, so there might be some bias given the indicators we chose. But, overall, we found that in order to predict GDP accurately, one should definitely use health expenditures as one of the major

factors, along with other major sectors we didn't cover in future work like employment rates, imports/exports.

Comparing the two models, we can see a consistent story with small differences. Both the linear regression and the random forest identify health factors as the primary correlates of GDP. The linear model gave us an understandable equation and confirmed the significance of health spending, life expectancy, literacy, and poverty, whereas the random forest reinforced health spending's dominance and allowed for non-linear insights. The linear model suggests that a one standard deviation improvement in health spending correlates with about a 0.8 SD higher GDP, and the random forest also implies countries in the top tier of health spending are much more likely to be in the top tier of GDP.

Conclusions and Discussion

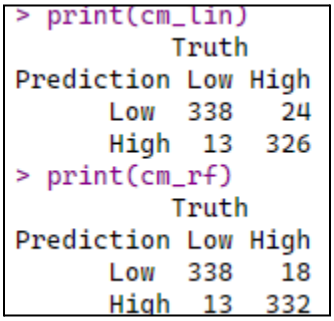

Figure 13: Confusion Matrix for Linear Regression (Top) and Random Forest (Bottom)

This project explored the linkages between economic wealth and human development indicators. Through an exploratory data analysis and the application of both linear and non-linear models like Linear Regression and Random Forest, we find that a country's GDP is highly predictable from a handful of health and education metrics. In particular, we saw that health expenditure per capita was the most notable and effective indicator in association with GDP. Countries that invest more in health tend to have significantly higher GDP per capita. This finding is consistent with the study from Becker's article, noting that health is a form of human capital, as healthier populations are more productive, can work longer, and focus on higher education (World Health Organization, 1). Education also shows a positive correlation with GDP, though in our analysis its effect was shown mainly through adult literacy rather than primary school enrollment. We can conclude from this that education alone is not enough to drive economic differences today's world. Instead, it's the quality and extent of education that differentiate high-income economies from the rest. The poverty rate being a significant negative predictor of GDP highlights how

poverty can hold back economic development. High poverty often means a large share of the population has low productivity, poor health, and limited access to education, which drags down overall per capita output.

Both the linear regression and random forest models point to a consistent policy message, which is to invest in health, sustain and improve education, and reduce poverty to promote economic growth. There was a lot of change that went on between each process, starting from the initial analysis, where we had to limit the range of data we had due to the large percentage of missing data. There was also results from the exploratory analysis that pointed at which indicators might be more helpful in prediction, which inspired changes and curiosity to test different combinations of indicators. In the end, all indicators were kept for modelling, just to make sure we highlight the bigger picture.

To discuss future work, it is worth first addressing the causal ambiguity, where richer countries can spend more on health and education. This brings the question that is it wealth enabling development, or development enabling wealth? The reality is likely a cycle where initial investments in health and education can kick-start growth, which then provides resources for further investment. Because of this, we need more factors in account other than human capital to full predict GDP per capita. It would be more accurate to include macroeconomic, governance, or infrastructure indicators like employment rates, corruption indices, and urbanization rates. Overall, human capital plays a enormous role on a country's development, more than what most will probably infer, but it is important to not also forget to keep other aspects of development into consideration as well.

References

Arel-Bundock, Vincent, and Etienne Bacher. *World Development Indicators and Other World Bank Data*. By World Bank, 2025, cran.r-project.org/web/packages/WDI/WDI.pdf.

Greenwell, Bradley Boehmke &. Brandon. *Chapter 11 Random Forests | Hands-On Machine Learning With R*. 1 Feb. 2020, bradleyboehmke.github.io/HOML/random-forest.html.

Health and the Economy. 6 May 2021, www.who.int/teams/health-financing-and-economics/economic-analysis/health-and-wealth#:~:text=Health%20systems%20are%20important%20contributors,health%20security%20and%20economic%20security.

"Human Capital - Econlib." *Econlib*, 27 July 2018, www.econlib.org/library/Enc/HumanCapital.html.

World Development Indicators | DataBank. databank.worldbank.org/source/world-development-indicators.