University rankings, economy, and human development

Andrea Valle & Adrian Imfeld

September 20, 2023

Contents

1 Introduction

In an increasingly interconnected and competitive global landscape, education has become a cornerstone of individual opportunity and national success. Universities serve as incubators for talent, research, and innovation, often becoming significant drivers of economic and human development. This data science project aims to investigate the relationships between university rankings, economic development, and human development by leveraging a multi-dimensional combined dataset spanning multiple years and countries.

University rankings, produced annually by several organizations, consider a range of metrics including research output, faculty qualifications, and global reputation. Economic development, frequently measured in Gross Domestic Product (GDP) per capita, encapsulates a nation's economic health and standard of living. Human Development Index (HDI), on the other hand, offers a composite score of life expectancy, education, gender equality, and income indices to provide a fuller understanding of human well-being.

The objective of this project is to unearth the hidden patterns and correlations among these variables over time, across a broad cross-section of countries. This project aims to reveal whether higher-ranked universities contribute directly or indirectly to economic and human development. The results of this research could have implications for educational policy, international development, and global talent migration, potentially serving as a guide to leverage education as a tool for holistic national advancement.

However, it's crucial to acknowledge the limitations of this small research project. One of the most significant limitations is that the project may identify correlations but cannot infer causation. In other words, even if the data suggests that countries with more and higher-ranked universities tend to have higher GDP per capita and better HDI scores, this does not prove that the universities themselves are the cause of these improvements. On the contrary, richer countries may simply be able to invest more into universities and education. Numerous confounding variables, such as government policies, cultural factors, and natural resources, could also contribute to these outcomes. The synergistic effects between these variables are complex and may involve lagging indicators or feedback loops that are beyond the scope of this project. Therefore, while the correlations unearthed may serve as a starting point for more targeted research, they should not be used as a basis for causal claims without further investigation.

2 Data preparation

2.1 Datasets

2.1.1 Academic ranking of world universities (Shanghai ranking)

According to their website, the Academic Ranking of World Universities (ARWU) is recognized as the precursor of global university rankings and the most trustworthy one. ARWU presents the world's top 1000 research universities annually based on transparent methodology and objective third-party data.

- Name: Academic Ranking of World Universities (ARWU) - Shanghai Rankings from 2005-2018
- Website: https://www.shanghairanking.com/
- Dataset: https://www.kaggle.com/datasets/joebeachcapital/shanghai-world-university-ranking

The dataset contains a row per university and year, assigning a rank to each university based on various indicators:

- Quality of Education, with Alumni and Award indicators (10% and 20% of the final mark)
- Quality of Faculty, with HiCi and N&S indicators (20% and 20% of the final mark)
- Research Output, with PUB indicator (20% of the final mark)
- Per Capita Performance, with PCP indicator (10% of the final mark)

These factors and their respective weights are combined to calculate the total score for each university, which is then used to rank them in the dataset. This ranking provides valuable insights into the overall performance and reputation of universities over time.

More on the methodology and definitions of the indicators can be found on their methodology website. The structure of the dataset looks like this (indicators and ISO country codes are ommitted):

World.rank	Rank	Year	University	Country
201-302	201	2009	University of Bochum	Germany
303-401	303	2009	The University of Texas Medical Branch at Galveston	United States
151-200	151	2010	Florida State University	United States
201-300	201	2010	Lancaster University	United Kingdom
101-150	101	2010	University of Groningen	Netherlands
301-400	301	2010	University of Southern Denmark	Denmark
16	16	2010	University of Washington	United States
151-200	151	2010	University of Hamburg	Germany
301-400	301	2010	Eotvos Lorand University	Hungary
71	71	2010	Ecole Normale Superieure - Paris	France

2.1.2 Economic dataset

To comprehensively assess a country's economic landscape, one of the most reliable indicators is the country's GDP. Specifically, it should account for population size to avoid overestimating the economic strength of larger nations. Therefore, we utilize the GDP per capita as a more accurate metric.

- Name: World GDP (GDP per capita)
- Dataset: https://www.kaggle.com/datasets/zgrcemta/world-gdpgdp-gdp-per-capita-and-annual-growths

The dataset consists of information for 206 countries, including their GDP per capita values from 1960 to 2020.

Country.Name	Code	X1990	X2020
Aruba	ABW	12306.7177	NA
Africa Eastern and Southern	AFE	696.9881	1359.6182
Afghanistan	AFG	NA	516.7479
Africa Western and Central	AFW	594.7260	1710.0734
Angola	AGO	948.3382	1776.1669
Albania	ALB	617.2304	5246.2923
Andorra	AND	18878.8523	NA
Arab World	ARB	2796.7000	5612.6866
United Arab Emirates	ARE	27729.3906	36284.5552
Argentina	ARG	4333.4834	8579.0178

2.1.3 Human Development Index (HDI)

The Human Development Index (HDI) was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The HDI is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living. The health dimension is assessed by life expectancy at birth, the education dimension is measured by mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The standard of living dimension is measured by gross national income per capita.

- Name: Human Development Index (HDI)
- Website: https://hdr.undp.org/data-center/human-development-index
- Dataset: https://hdr.undp.org/sites/default/files/2021-22_HDR/HDR21-22_Composite_indices_complete_time_series.csv

- Metadata: https://hdr.undp.org/sites/default/files/2021-22_HDR/HDR21-22_Composite_indices_metadata.xlsx

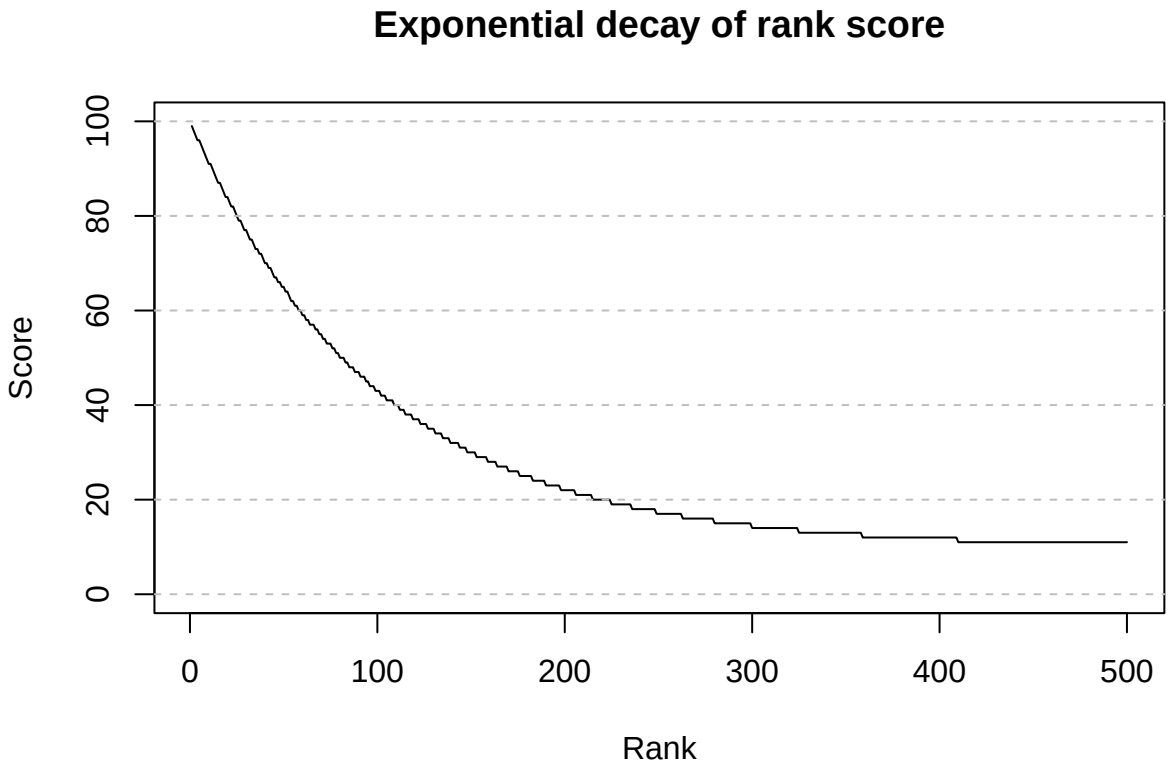
The dataset is in wide format, containing a row for each of the 206 countries and a column for each indicator-year combination. The time span ranges from 1990-2021 for most indicators, resulting in 1008 columns total. The structure looks like this (only a selected few indicator-year columns are shown):

iso3	country	hdicode	hdi_1990	hdi_2021	le_1990	le_2021
AFG	Afghanistan	Low	0.273	0.478	45.9672	61.9824
AGO	Angola	Medium	NA	0.586	41.8933	61.6434
ALB	Albania	High	0.647	0.796	73.1439	76.4626
AND	Andorra	Very High	NA	0.858	78.4063	80.3684
ARE	United Arab Emirates	Very High	0.728	0.911	71.9004	78.7104
ARG	Argentina	Very High	0.723	0.842	71.7837	75.3899
ARM	Armenia	High	0.656	0.759	68.8209	72.0431
ATG	Antigua and Barbuda	High	NA	0.788	73.4922	78.4968
AUS	Australia	Very High	0.865	0.951	77.0443	84.5265
AUT	Austria	Very High	0.825	0.916	75.7396	81.5797

2.2 Preprocessing

2.2.1 Univerity rank aggregation

In order to correlate the university rankings with economic and human development over time, the university rankings were aggregated per country and year. First, a score was assigned per university based on its rank according to an exponential decay model. The idea is to score top universities a lot higher and introduce an exponential decay in score for lower ranks. This approach was inspired by webometrics, which orders countries by the number of universities in top categories first.



Once every university was assigned a score per year based on its rank, the scores were summed up and the universities counted per country and year. The aggregation resulted in the following structure, ready to be joined with the economic and HDI datasets:

ISO3.CODE	Country	Year	Points	Count
USA	United States	2005	6448	168
USA	United States	2006	6366	167
USA	United States	2007	6373	166
USA	United States	2008	6289	159
USA	United States	2009	6186	152

2.2.2 Economic dataset

To pursue our research goals, we intend to refine the GDP dataset to include the years from 2005 to 2018, aligning with the time coverage of the university ranking data. After this adjustment, we will proceed to restructure the dataset into a long format to facilitate our analytical efforts.

Country.Name	Code	Year	gdp.per_capita
United States	USA	2005	44114.75
United States	USA	2006	46298.73
United States	USA	2007	47975.97
United States	USA	2008	48382.56
United States	USA	2009	47099.98

2.2.3 HDI dataset

From the Human Development Index (HDI) dataset, the timelines of the following indicators of interest were picked:

- Human Development Index – a summary of a long and healthy life, access to knowledge and a decent standard of living.
- Life expectancy at birth
- Expected years of schooling
- Mean years of schooling
- Gross National Income (GNI) per capita
- Gender Development Index (GDI) – measures gender inequalities in achievement in three basic dimensions of human development:
  - health, measured by female and male life expectancy at birth
  - education, measured by female and male expected years of schooling for children and female and male mean years of schooling for adults ages 25 years and older
  - command over economic resources, measured by female and male estimated earned income
- Gender Inequality Index (GII) – reflects gender-based disadvantage in three dimensions for as many countries as data of reasonable quality allow:
  - reproductive health
  - empowerment
  - labourmarket

For details on the indicators, see the technical notes.

The dataset was transformed into a suitable format, containing a row per country and year, and columns for the indicators. After joining the transformed HDI dataset with the previously prepared university ranking data, world region and continent information was joined as well as country population data from the world bank dataset. This is the structure of the final dataset (some indicators are omitted):

country	continent	year	population	hdi	le	eys	mys	gnipc	uni_score	uni_count
Germany	Europe	2005	82469422	0.91	79.26	16.34	13.70	44193.77	1059	40
Germany	Europe	2006	82376451	0.92	79.59	16.39	13.60	46045.61	1047	40
Germany	Europe	2007	82266372	0.92	79.74	16.44	13.68	47424.30	1049	41
Germany	Europe	2008	82110097	0.92	79.83	16.48	13.76	47413.82	1053	40
Germany	Europe	2009	81902307	0.92	79.92	16.53	13.83	46040.43	1056	40

2.3 Merge datasets

Having all three core datasets available, it is now appropriate to combine them into one complete dataset.

Our first step was to merge the pre-existing dataset with the human development dataset. Subsequently, it has been followed a similar process to integrate the university rankings and GDP per capita datasets, using country codes and years as the key merging parameters.

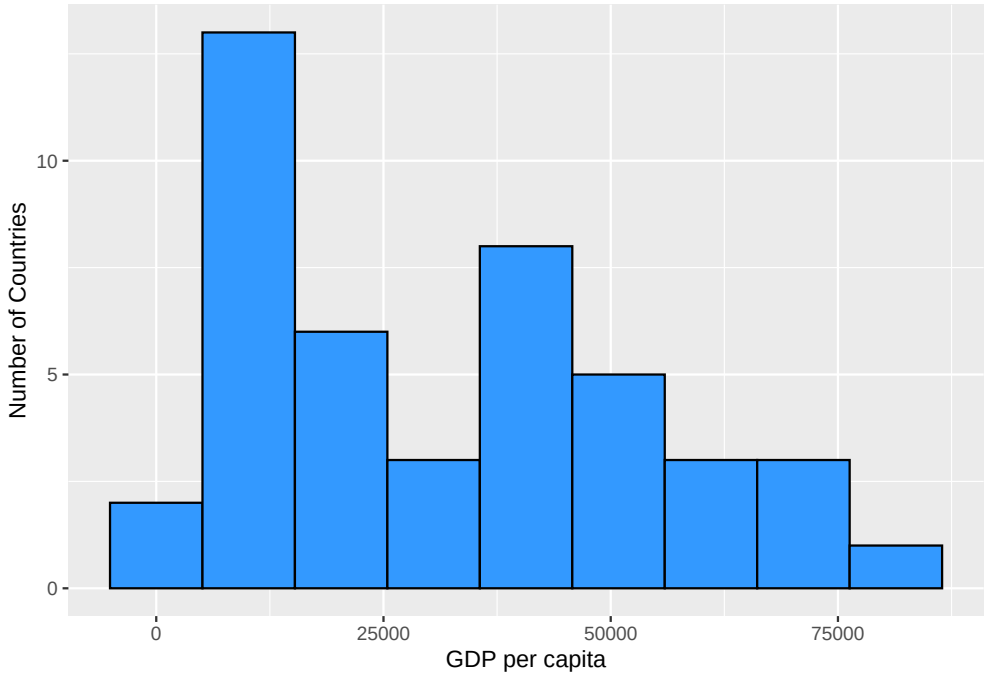
country	University	Year	hdi	le	mys	gnipc	uni_score	uni_count	gdp.per_capita
Germany	University of Bochum	2009	0.92	79.92	13.83	46040.43	1056	40	41485.90
Germany	University of Hamburg	2010	0.93	80.09	13.85	47554.23	1020	39	41531.93
Germany	University of Duisburg-Essen	2010	0.93	80.09	13.85	47554.23	1020	39	41531.93
Germany	University of Frankfurt	2015	0.94	80.57	14.08	52200.63	967	39	41086.73
Germany	University of Halle-Wittenberg	2015	0.94	80.57	14.08	52200.63	967	39	41086.73

3 Economic analysis

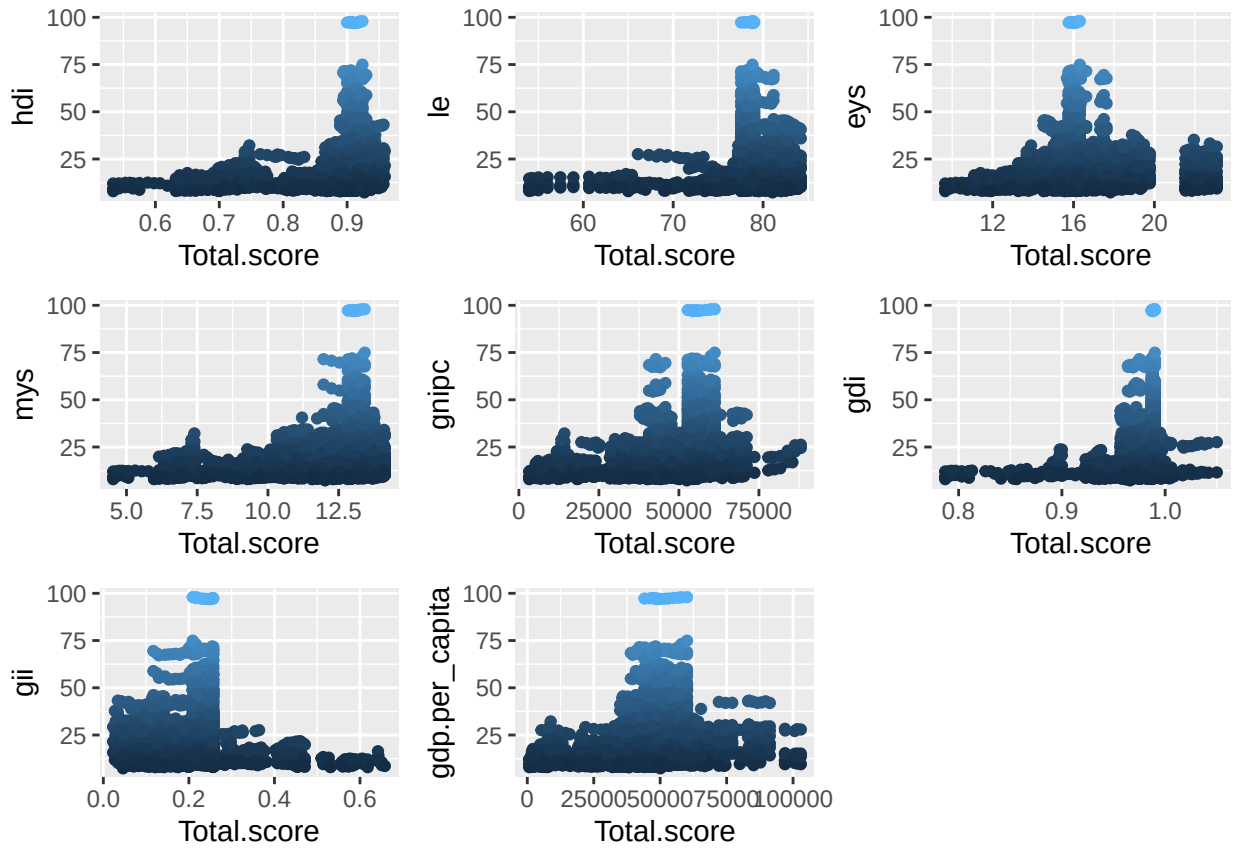
3.1 Exploratory analysis

In this section, an exploration of the intricate connection between education, HDI scores and GDP per capita will be conducted. The focus is on examining the various factors influencing this relationship and shedding light on the significance of education as a driver of economic growth and societal advancement. The objective is to gain a deeper understanding of how a nation's economic well-being and human development factors affect university rankings.

To gain an initial understanding of GDP per capita, it is crucial to analyze its distribution, which exhibits a noticeable right-skew. This skewness suggests that most countries in the dataset have lower GDP per capita values, highlighting the prevalence of lower income levels among the nations included.



The exploratory analysis progresses by scrutinizing the factors that impact the total points. The aim of this analysis is to detect potential predictors and assess the connections between variables. To begin this exploration, a series of variables were plotted alongside the total score.



3.2 Modeling Total Score

Utilizing correlation analysis, variables presenting non-linear patterns were identified. Subsequently, a Generalized Additive Model (GAM) was chosen to build the model, incorporating the following variables: Total.score, hdi, le, eys, mys, gnipc, gdi, gii, uni_score, gdp.per_capita.

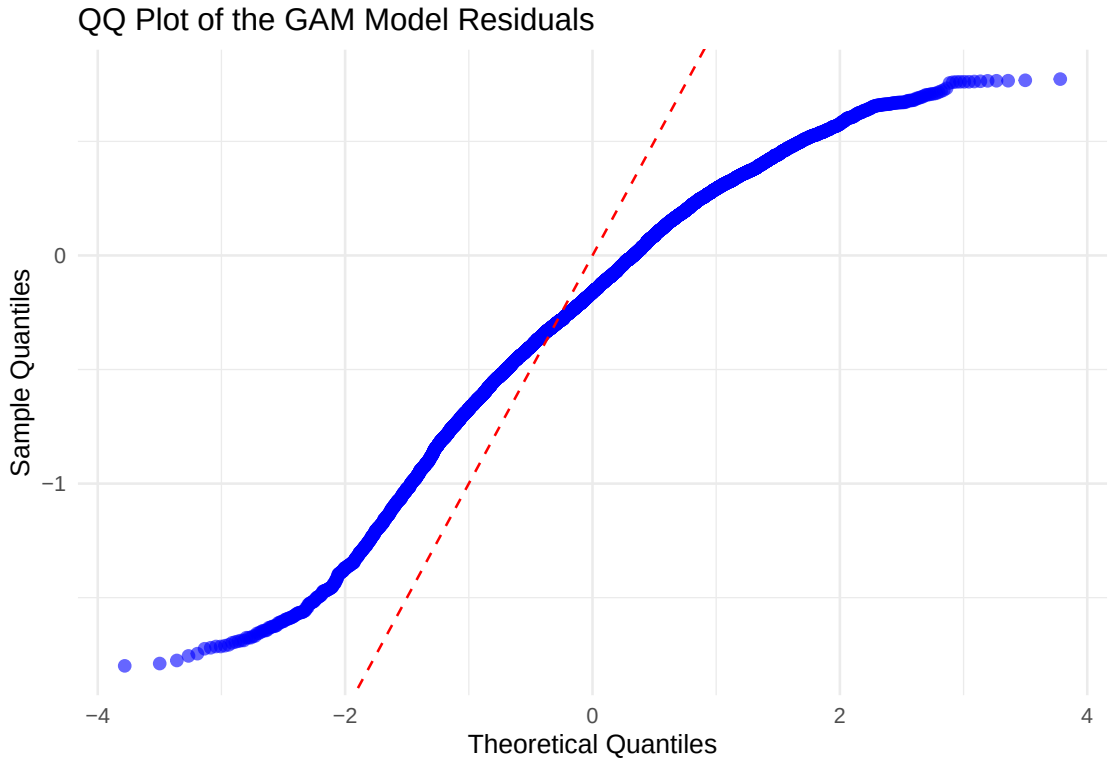
	edf	Ref.df	F	p-value
s(gdp.per_capita)	2.403497	3.121490	1.0396003	0.4189473
s(hdi)	7.175975	8.021461	1.9405381	0.0532418
s(le)	1.008524	1.014583	0.4785134	0.4916571
s(eys)	4.961207	6.147043	2.0520244	0.0500711
s(mys)	7.871809	8.437914	6.5545831	0.0000000
s(gnipc)	5.408243	6.490635	1.0780034	0.4162458
s(gdi)	6.385555	7.324214	3.0920168	0.0036850
s(gii)	5.656513	6.636512	5.3596265	0.0000054

In our analysis, it is evident that three specific variables, namely "Average Years of Schooling" (mys), "Gender Inequality Index" (gii) and "Gender Development Index" (gdi), were found to be statistically significant in the context of the GAM model we are trying to fit. These results indicate that these variables exert substantial influence on the outcome of our model and that their contributions are statistically significant.

Collectively, these results underscore the importance of mys and GII as influential factors contributing to the variability of our response variable. They constitute essential components of our model, enriching its ability to capture and explain the underlying patterns in the data.

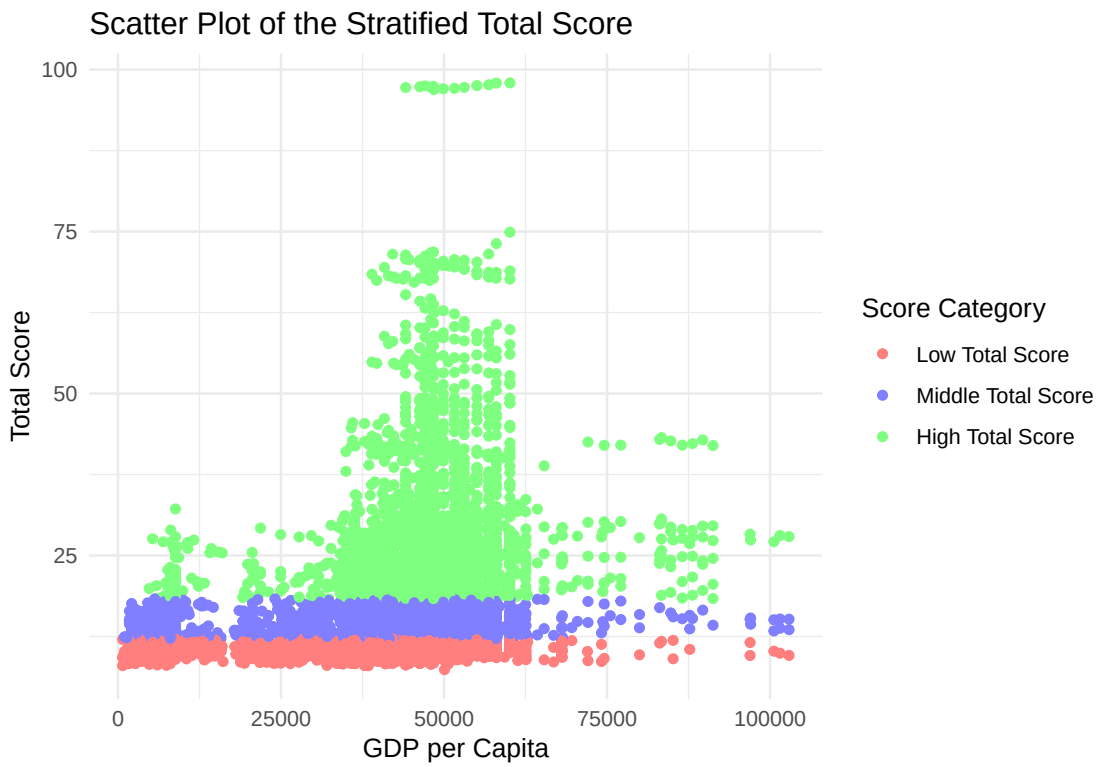
These findings collectively underscore the importance of mys, GII, and GDI as influential factors contributing to the variability of our response variable. They constitute essential components of our model, enriching its ability to capture and explain the underlying patterns in the data. Furthermore, the presence of GII and GDI in the model emphasizes the significance of addressing gender-based disparities and promoting gender equity as integral components of broader development efforts. A higher GII value reflects greater gender inequality, suggesting that reducing such disparities is associated with improvements in the Total Score. Similarly, a higher GDI indicates better gender development outcomes, highlighting the role of gender equity as a contributing factor to overall development.

To analyze the results of the modeling, it is beneficial to closely examine the QQ plot of the residuals, as it offers insights into the distribution of model errors.

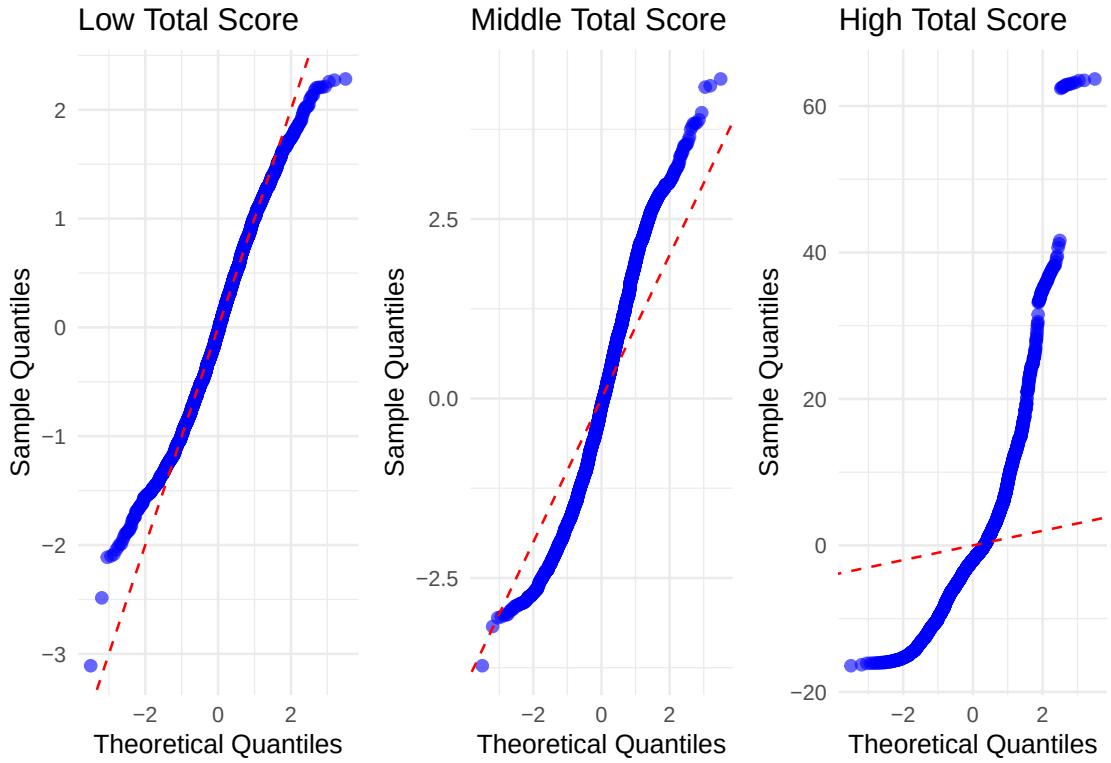


In the QQ plot of the residuals for the Generalized Additive Model (GAM), we can observe that the residuals tend to deviate slightly below the theoretical normal line. This behavior suggests a minor departure from the normal distribution, indicating that the GAM model may not perfectly capture the distribution of the residuals. While the deviations are relatively small, it's important to note that the normality assumption of the residuals may not be fully met. Further investigation and potentially considering alternative models could be beneficial to improve the model fit and better capture the data distribution.

To enhance the goodness of fit of the generalized additive model (GAM), a stratification strategy is employed. This strategy involves dividing the entire dataset into three distinct categories based on quantile divisions of the Total Score. It's important to note that this stratification encompasses all observations across all years simultaneously, without any year-based filtering. The primary objective of this approach is to improve overall model performance and enhance the accuracy of estimates.



To analyze the results of the modeling, it is beneficial to closely examine the QQ plot of the residuals, as it offers insights into the distribution of model errors.



The QQ plots of the residuals for the three income categories do not reveal any substantial improvement in

the goodness of fit of the GAM model. It suggests that further investigation is required to achieve a more accurate fit to the data. An intriguing observation is that the model's accuracy appears to diminish as the total score increases.

3.3 Modeling Score based on the exponential decay

Following the preprocessing steps conducted initially, a scoring system was implemented for universities based on their rankings, utilizing an exponential decay model. The objective was to assign significantly higher scores to top-ranked universities while introducing a gradual decrease in score for lower-ranked institutions. The focus shifted from the total score to the newly assigned scores to evaluate how effectively the GAM model fits this updated data.

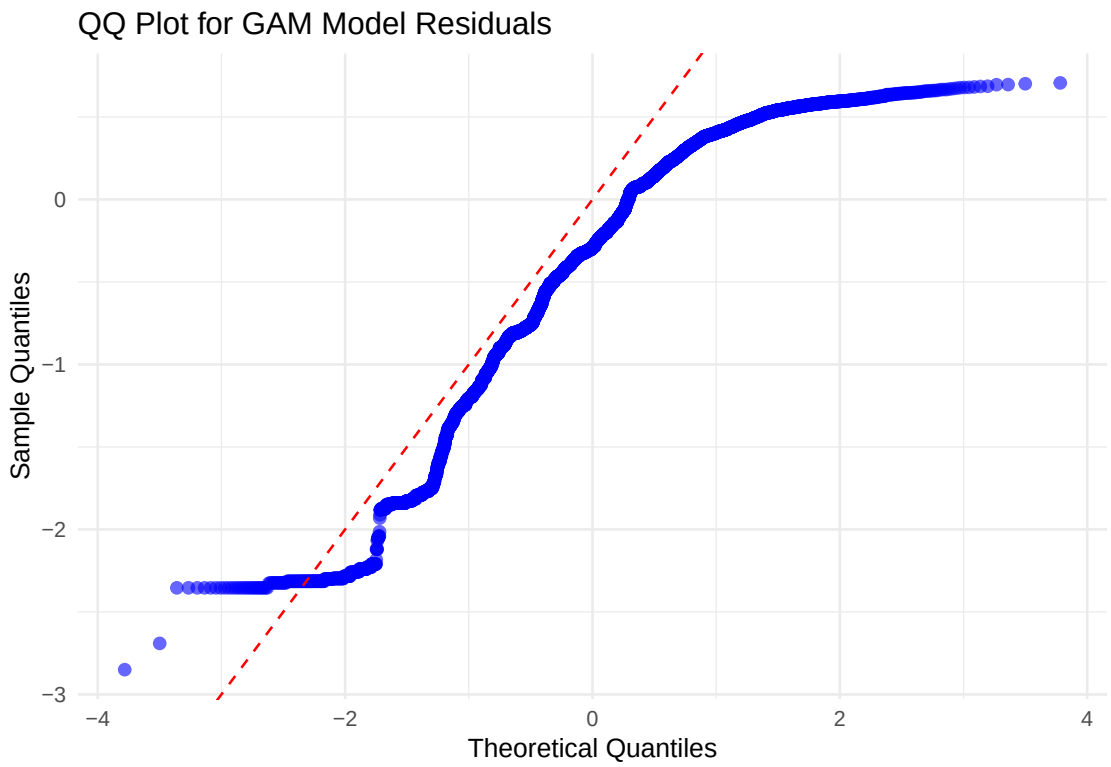
	edf	Ref.df	F	p-value
s(gdp.per_capita)	2.070137	2.721585	0.6227110	0.5054827
s(hdi)	8.011730	8.602531	2.1362997	0.0789250
s(le)	1.001184	1.002139	1.8768375	0.1704689
s(eys)	1.000212	1.000406	0.7393212	0.3896615
s(mys)	7.331141	8.099918	7.5034353	0.0000000
s(gnipc)	8.537144	8.903935	3.6278389	0.0001502
s(gdi)	6.234348	7.203503	4.4463717	0.0000862
s(gii)	6.165777	7.177551	7.4128890	0.0000000

In the context of this analysis, several socio-economic and developmental indicators have emerged as statistically significant contributors to the Generalized Additive Model (GAM) under examination. These variables collectively add complexity to the GAM, with each making a distinctive and statistically significant impact on the model's ability to understand and elucidate the observed variability within the response variable.

Among these variables, "Average Years of Schooling" (mys) underscores the pivotal role of education in human development, enhancing the model's predictive capabilities. Additionally, "Gross National Income (GNI) per capita" (gnipc) emphasizes the critical importance of economic well-being within the broader context of overall development.

Furthermore, the "Gender Development Index" (GDI) indicates that nations with equitable gender development tend to achieve higher Total Scores, underscoring the significance of gender equity in contributing to overall development. Similarly, the "Gender Inequality Index" (GII) highlights the need to address and reduce gender-based inequalities, given its substantial influence over Total Scores. These findings collectively underscore the intricate interplay between socio-economic and developmental factors and the response variable.

These variables collectively contribute to the complexity of our GAM, with each making a unique and statistically significant impact on the model's ability to capture and explain the variability in our response variable. Their inclusion underscores the multidimensional nature of human development and its intricate interplay with the response variable under investigation.

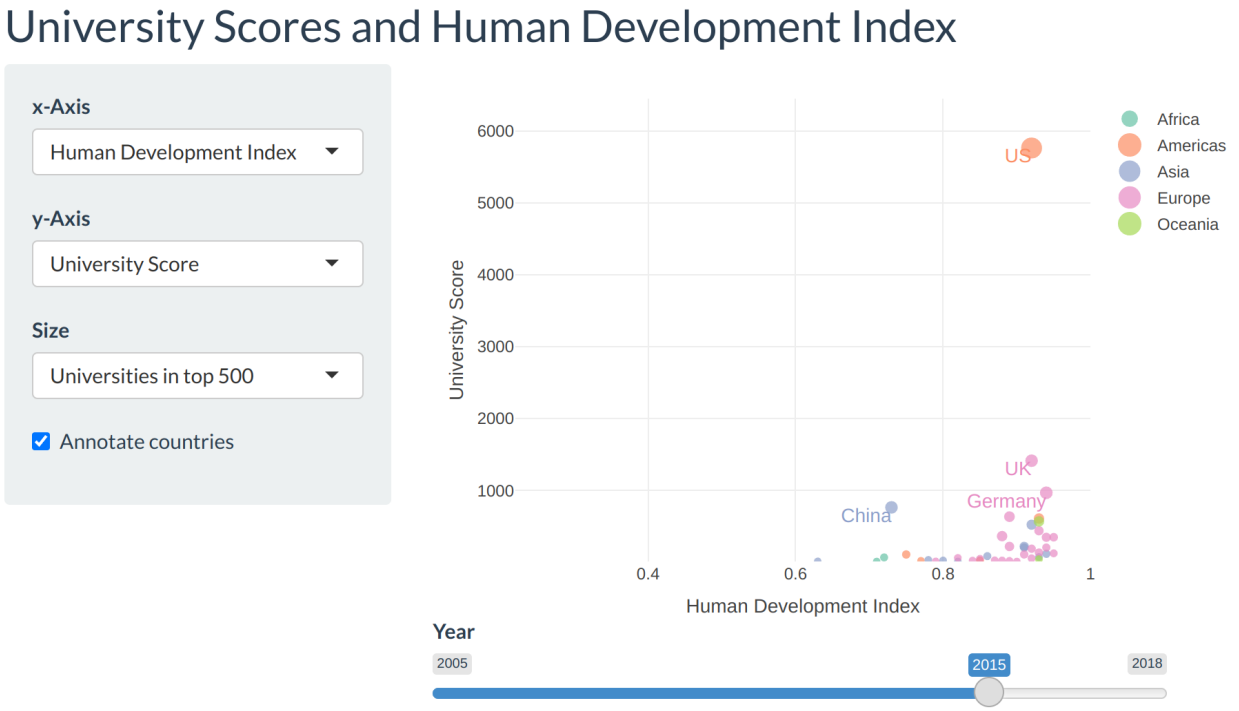


The QQ plot analysis, when applied to the "Points" variable instead of the "Total Score," does not yield a significant improvement. This suggests that switching from "Total Score" to "Points" does not result in a better fit of the Generalized Additive Model (GAM) to the data.

4 Human development

4.1 Interactive app

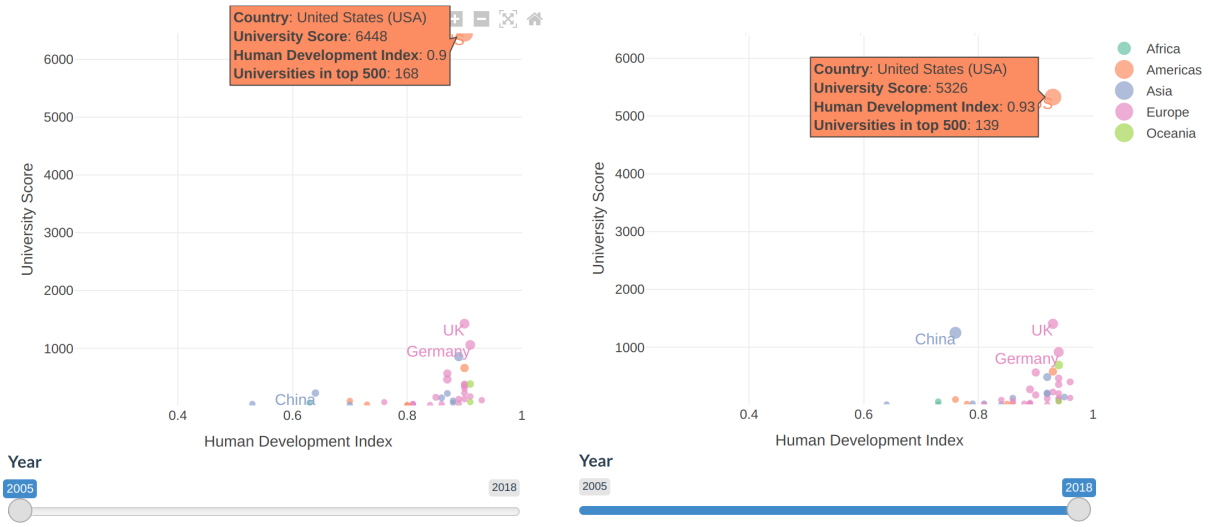
In order to explore the relationships between the university rankings and the HDI indicators, an interactive R shiny app was built and published here: https://aimfeld.shinyapps.io/uni-ranks. The source code and data is available on Github.



The user interface allows for selecting an indicator for the x-axis, y-axis, and marker size. Development over time can be observed by dragging the year slider. Finally, some countries of interest annotations can be toggled on and off. Since the university ranking data only spans the years 2005-2018, the time range is adjusted dynamically when selecting the indicators.

The app is inspired by the famous Gapminder animations, presented by Hans Rosling in a very entertaining TED talk.

4.2 University scores and human development

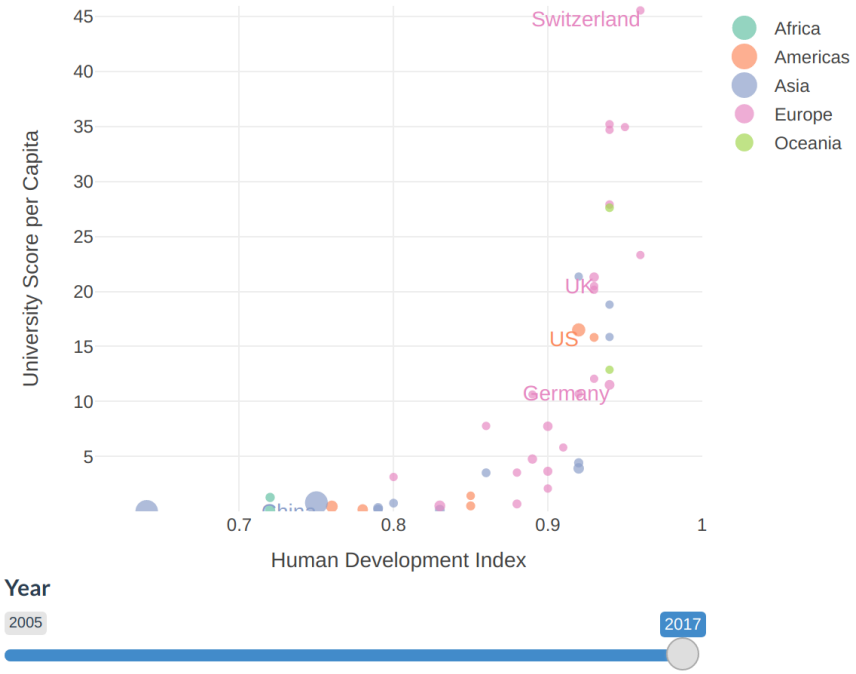


In terms of university scores, the US was dominating with 168 universities in the top 500 in 2005. Over the years, China has been catching up in number and ranking of top 500 universities as well as in human development. The decline in combined university scores of the US from 2005 to 2018 may be explained by Chinese universities entering the top 500 and displacing some of the lower ranked US universities. However,

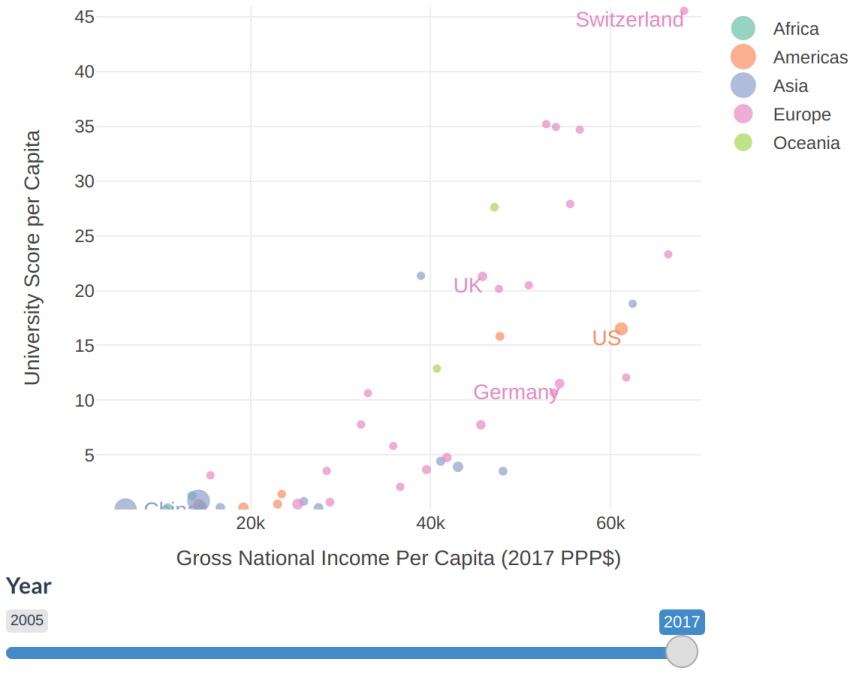
US dominance remains very strong. In 2023, the top ranks are is still heavily dominated by US universities. In fact, according to the Shanghai rankings 2023, 20 of the top 30 are US universities whereas only 2 are Chinese. These are the top 10 universities in 2023:

World Rank	Institution	All	National/Regional Rank	Total Score	Alumni
1	Harvard University		1	100.0	100.0
2	Stanford University		2	74.8	45.6
3	Massachusetts Institute of Technology (MIT)		3	69.1	72.5
4	University of Cambridge		1	67.9	78.9
5	University of California, Berkeley		4	63.4	65.8
6	Princeton University		5	60.1	61.7
7	University of Oxford		2	59.5	49.9
8	Columbia University		6	55.3	58.5
9	California Institute of Technology		7	54.5	56.1
10	University of Chicago		8	53.8	56.6

The relationship between university scores and HDI is unclear. However, an exponential relationship emerges if we look at university score *per capita*, correcting for country population size. Notably, Switzerland comes out on top due to its small country size and comparatively high number of quality universities.



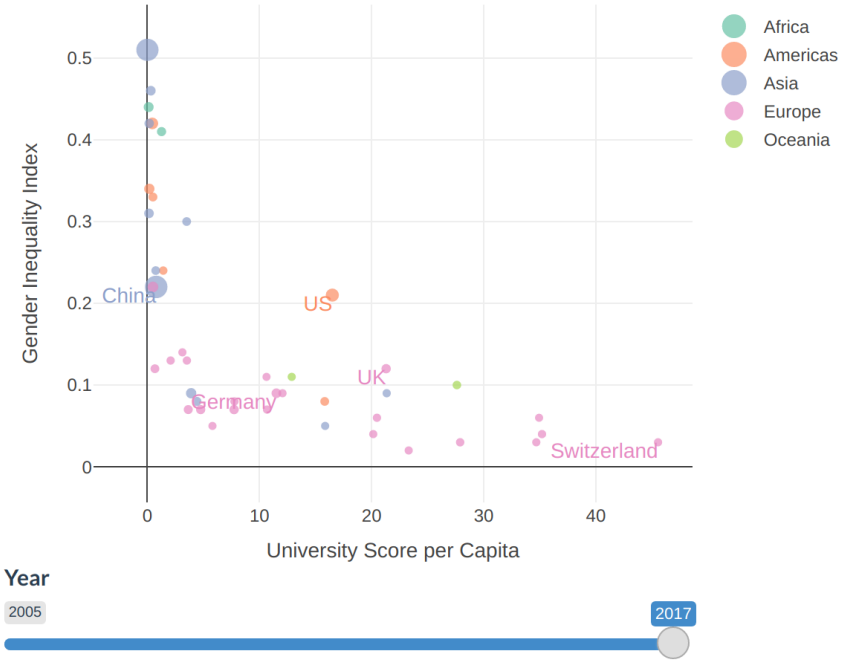
A very similar exponential pattern can be found between Gross National Income (GNI) and university score per capita. This is not surprising, since the GNI is part of the HDI.



Countries may start investing heavily into education, once a certain level of human development and wealth is reached which can lead to a positive feedback loop. Better education may in turn drive human development and GNI, which could explain the observed exponential relationship.

4.3 Education and gender inequality

We might hypothesize that better education leads to less gender inequality. Indeed, the higher the university score per capita, the lower the gender inequality index. Asian countries seem to lag behind in terms of gender equality and education, while European countries are doing quite well. However, the situation has been improving in Asia over time, as can be observed interactively by using the year slider in the app.



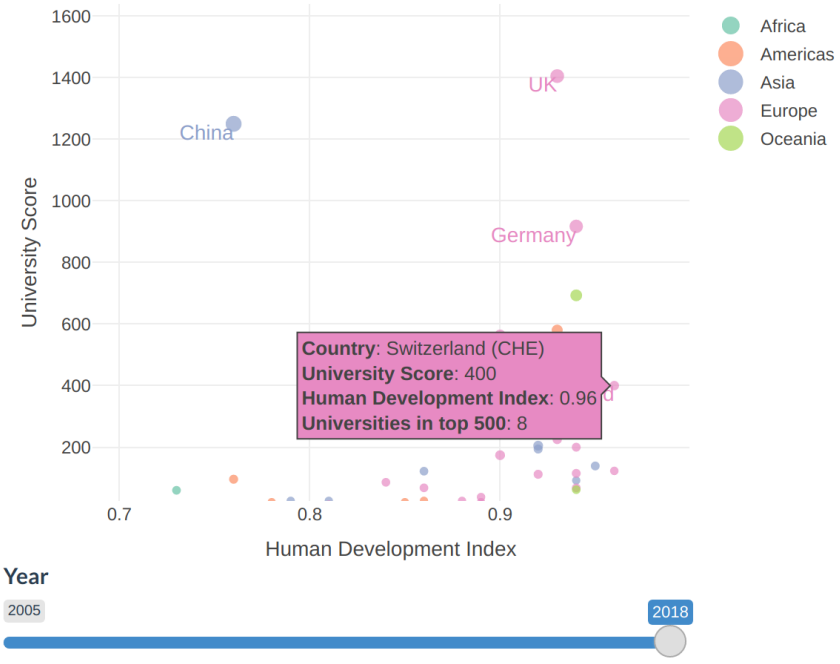
While it's plausible to infer a causal relationship between education and gender equality, both of these variables may be mediated by hidden variables such as economic development or other factors.

4.4 Switzerland

For being a very small country, Switzerland is doing very well in terms of top ranking universities and human development. There are consistently around 8 Swiss universities in the top 500. In 2023, the top Swiss universities include the following, spearheaded by the ETH Zürich with a rank of 20.

World Rank	Institution	Switzerland	National/Regional Rank	Total Score	Alumni
20	ETH Zurich		1	44.1	26.7
49	University of Geneva		2	31.8	37.7
54	Swiss Federal Institute of Technology Lausanne		3	31.3	0.0
59	University of Zurich		4	30.7	0.0
81	University of Basel		5	26.8	21.6
101-150	University of Bern		6-7		11.7
101-150	University of Lausanne		6-7		16.6

With respect to human development, Switzerland is equalled only by Norway with a HDI of 0.96.



5 Conclusion

In summary, the generalized additive model (GAM) proved its reliability and adaptability, particularly when it used "points" as the target variable. This choice allowed our model to accommodate a wider range of predictor variables, thus improving its ability to capture and decipher the complexities inherent in the dataset. As a result, GAM emerges as the most reliable and robust modeling technique capable of accommodating a wide range of variables, thus providing valuable insights for our analysis. This experience can be leveraged to improve score predictions through the use of socioeconomic and developmental indicators.

In the context of our modeling efforts for both Total Score and Score the various socio-economic and developmental indicators emerged as statistically significant contributors to the Generalized Additive Model (GAM) under consideration. These variables collectively contribute to the complexity of our GAM, with a distinctive and statistically significant impact on the model's ability to understand the variability observed in the respective response variables.

In particular, "Average Years of Schooling" (mys) emphasizes the central role of education in human development, enhancing the model's predictive capabilities. In addition, "Gross National Income per Capita" (Gnipc) emphasizes the critical importance of economic well-being in the broader context of overall development.

In addition, the "Gender Development Index" (GDI) indicates that nations with equitable gender development tend to have higher total scores, underscoring the importance of gender equity in contributing to overall development. Similarly, the "Gender Inequality Index" (GII) highlights the need to address and reduce gender-based inequality, given its substantial influence on Total Scores. Taken together, these results underscore the intricate interplay between socioeconomic and development factors and their respective response variables, shedding light on the multifaceted nature of human development.