Exploring Global Education and Its Contributing Factors

by Sehajroop Singh and Marissa Santiago

Abstract

Due to tertiary education becoming increasingly valuable and playing an important role in a persons social and economic success, it is important to observe the factors that assist in making higher education more attainable. This project aims to study the relationship between a child's attained education and 8 external factors. After conducting exploratory analyses, we have discovered that major contributors to education success include daily maximum temperature of your country, average income and education exepnditure per inhabitant, along with Gender Equality.

Introduction

Background

Nearly 60% of all jobs in the U.S economy require higher education. The wage gap between people who have bachelor's degrees and people with only a high school diploma has nearly doubled since the early 1980s. However, how do these numbers match up to other scholars in other countries? has been an increase in peoples educations in the past years; 220 million tertiary education students in the world, up from 100 million in 2000.

Tertiary education has become increasingly valuable and plays an important role in our social and economic success. Tertiary education refers to all formal post-secondary education, including public and private universities, colleges, technical training institutes, and vocational schools. It is instrumental in fostering growth, reducing poverty, and boosting shared prosperity. Therefore it is important to observe the factors that assist in making higher education more attainable. However it can be challenging to determine the social and economic factors that influence a person's education level and how these factors vary across countries.

Our data is collected for a variety of organizations and assembled by the World Bank. The world bank group are five institutitons that work together to provide funding and knowledge for developing countries. The data for this project is a merge of The Global Database on Intergenerational Mobility (GDIM), EqualityData, and Politicalstability dataset. Findings in our data suggest regional location and educational expenditure per resident have a large impact on a person's educational success and highlights the factors that may influence educational growth.

Aims

This project aims to study the correlation between the total years spent in education for children and other external factors such as average parental education, countries of residence: education expenditure, average income, gender equality, political stability and mean temperature.

Data description

The data for this project is a merge of The Global Database on Intergenerational Mobility (GDIM), EqualityData, and Politicalstability dataset.

The infomation in the GDIM comes from a variety of surveys offered within each country between 2008 and 2016. This database includes educational mobility estimates from 153 economies and accounts for 97 percent of the world's population. Except for the Middle East and North Africa, the population coverage in all regions exceeds 90 percent, whilst in the middle east and North africa 83 percent of the population is covered.

The Gender Equality data set provides a few key variables, namely the Gender equality Index GEI, which is an aggregate of a few of the subsets of its variables. The World Economic Forum compiles and releases the Global Gender Gap Index every year. This report measures the extent of gender-based gaps among four key dimensions: Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment, then gives each country a ranking between 0.000 (or 0%, the lowest possible gender equality) and 1.000 (100%, the highest possible gender equality). The analyses of each country are intended to serve as a basis for designing effective measures for reducing gender gaps. The 2021 edition of the Global Gender Gap Index studied and ranked 156 countries and territories around the world.

The IQ by country dataset contains information about the average IQ of a country along with average income in US dollars and education expenditure per individual. The displayed IQ was averaged out of the results of 9 international studies and compared to the average income and government expenditures on education for the years 1990 to 2010.

The Worldwide Governance Indicators (WGI) are a research dataset summarizing the views on the quality of governance provided by a large number of enterprise, citizen and expert survey respondents in industrial and developing countries. These data are gathered from a number of survey institutes, think tanks, non-governmental organizations, international organizations, and private sector firms. The WGI does not reflect the official views of the World Bank, its Executive Directors, or the countries they represent. The WGI is not used by the World Bank Group to allocate resources." The regional code dataset was used to provide a variable with more specific sub-region locations for each country in order to better facilitate our understanding of how different parts of the world are affected by the other variables.

The regional code dataset was used to provide a variable with more specific sub-region locations for each country in order to better facilitate our understanding of how different parts of the world are affected by the other variables.

The data is publicly available:

> https://datacatalog.worldbank.org/search/dataset/0050771/global-database-on-intergenerational-mobility
>
> https://www.worlddata.info/iq-by-country.php

> https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv
>
> https://worldpopulationreview.com/country-rankings/gender-equality-by-country
>
> http://info.worldbank.org/governance/wgi/

Sample and measurement information

Data structure

For this study, the observational units is educational mobility in a given country. Education levels are recorded for both parent and child, comparing information on how child mobility changes by parent and region.

> **Table 1**: variable descriptions and units for each variable in the dataset.

Name	Variable description	Type	Units of measurement
country	Name of country	categorical	none
region	Region (with high-income economies among the regions)	categorical	none
sub-region	designated sub-region of region	catagorical	none
incgroup4	Income groups (4 categories) as of July 1, 2020	categorical	none
year	Survey year	categorical	none
parent	Mothers/Fathers/Max/Average	categorical	none
child	Sons/Daughters/All	categorical	none
MEANp	Mean of parents' years of schooling	numeric	none
MEANc	Mean of children's years of schooling	numeric	none
SDp	Standard deviation of parents' years of education	numeric	none
SDc	Standard deviation of children’s years of education	numeric	none
CAT	is the share of respondents that have attained a higher educational category than their parents, conditional on the parents not having obtained tertiary education, such that all included individuals have a chance of surpassing their parents. For this measure, we categorize individuals and parents according to their highest educational attainment in the following categories (see	numeric	range 0-1

	Section 4 for more details): (i) less than primary, (ii) primary, (iii) lower-secondary, (iv) upper-secondary, or (v) tertiary.		
GEI %	Gender equality index	numeric	range 0-1
Gender Economic Participation and Opportunity Equality %	Economic opportunities to participate in for gender	numeric	range 0-1
Gender Education Attainment Equality %	educational attainment for gender	numeric	range 0-1
Gender Health and Survival Equality %	Health and survival equality for gender	numeric	range 0-1
Gender Political Empowerment	Political empoerment rate of equality for gender	numeric	range 0-1
Political Instability Estimate	political instability rating of a given country	numeric	none
IQ	Average IQ of a country	numeric	none
Average Income (USD)	average income of residents in a country	numeric	USD Dollar
Education expenditure\nper inhabitant (USD)	average expenditure spent by inhibitants of a country	numeric	Dollar
Daily maximum\ntemperature Celsius	average daily temperature in given region	numeric	Celcius - C

In our data preprocessing stage,the initial dataset did not include details on on Gender equality, government instability, IQ, income, and temperature. Since, this additional data information would help with our analysis, we added the additional information by merging EqualityData and Politicalstability dataset with our GDIM dataset grouping by and country and sub-region. The variables included in the final dataset are displayed in Table 1 while the first few rows of the final dataset are shown in Table 2.

> **Table 2**: Example head rows and columns from the final GDIM dataset.

	country	region	incgroup4	parent	child	MEANp	MEANc	SDp	SDc
0	Afghanistan	South Asia	Low income	avg	all	2.480162	5.582401	3.809782	5.8577
1	Afghanistan	South Asia	Low income	avg	daughter	2.989992	4.331731	4.010379	5.9026
2	Afghanistan	South Asia	Low income	avg	son	2.353979	5.891941	3.748495	5.8061
3	Afghanistan	South Asia	Low income	dad	all	3.756049	5.538275	5.493762	5.8420
4	Afghanistan	South Asia	Low income	dad	daughter	4.520877	4.277768	5.804775	5.8697

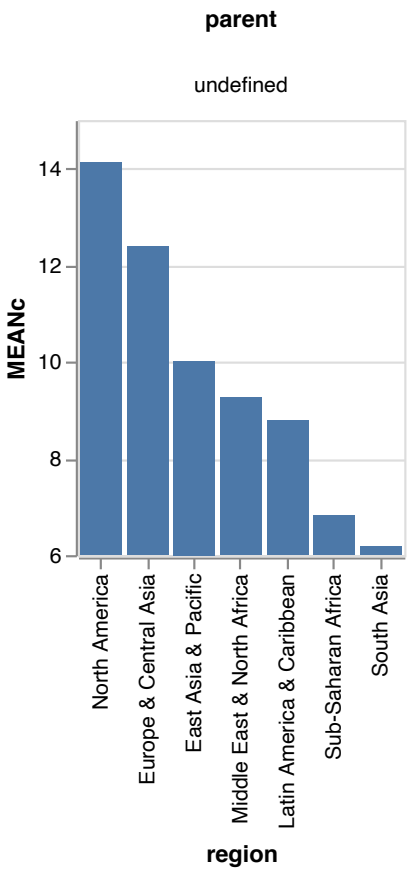
Methods

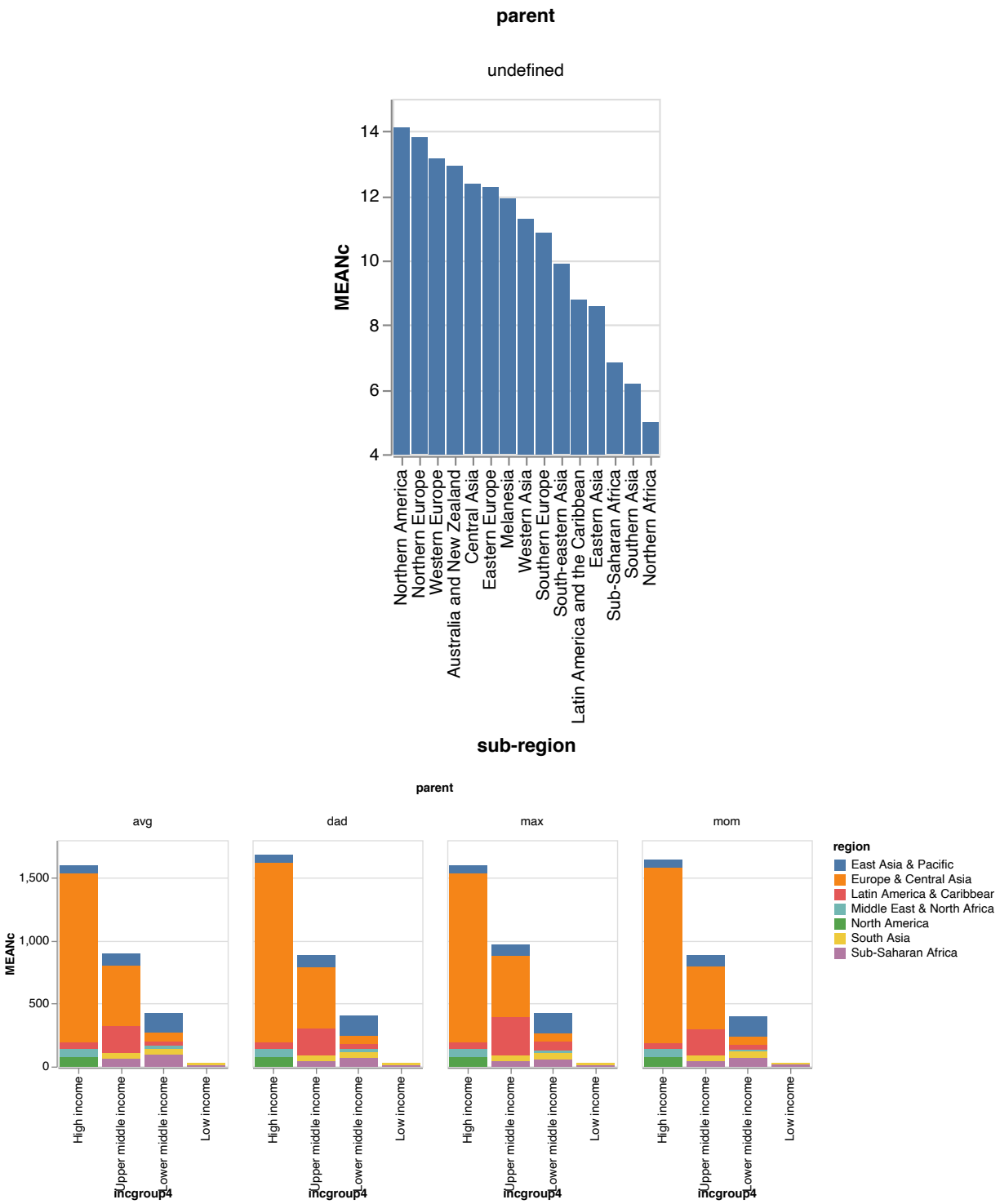
Exploratory analysis aimed at illuminating any clear correlations between Education and our set of variables . At this stage of our analysis we checked the relationship visually to see what kind of model structure is sensible. Subsequently, principal components analysis was performed on the normalized GDIM dataset to identify measures of intergenerational mobility that capture a significant portion of total variation in the data; typical values of these measures were compared by region. We then performed a regression analysis to see how strongly each variable effect average education of the child.

Results

Education Exploration

> **Figure 1**: Histogram (Left) MEANc within regions, Histogram (Center) MEANc within subregions, Histogram (Right) MEANc whe related to MEANp distinguished by region



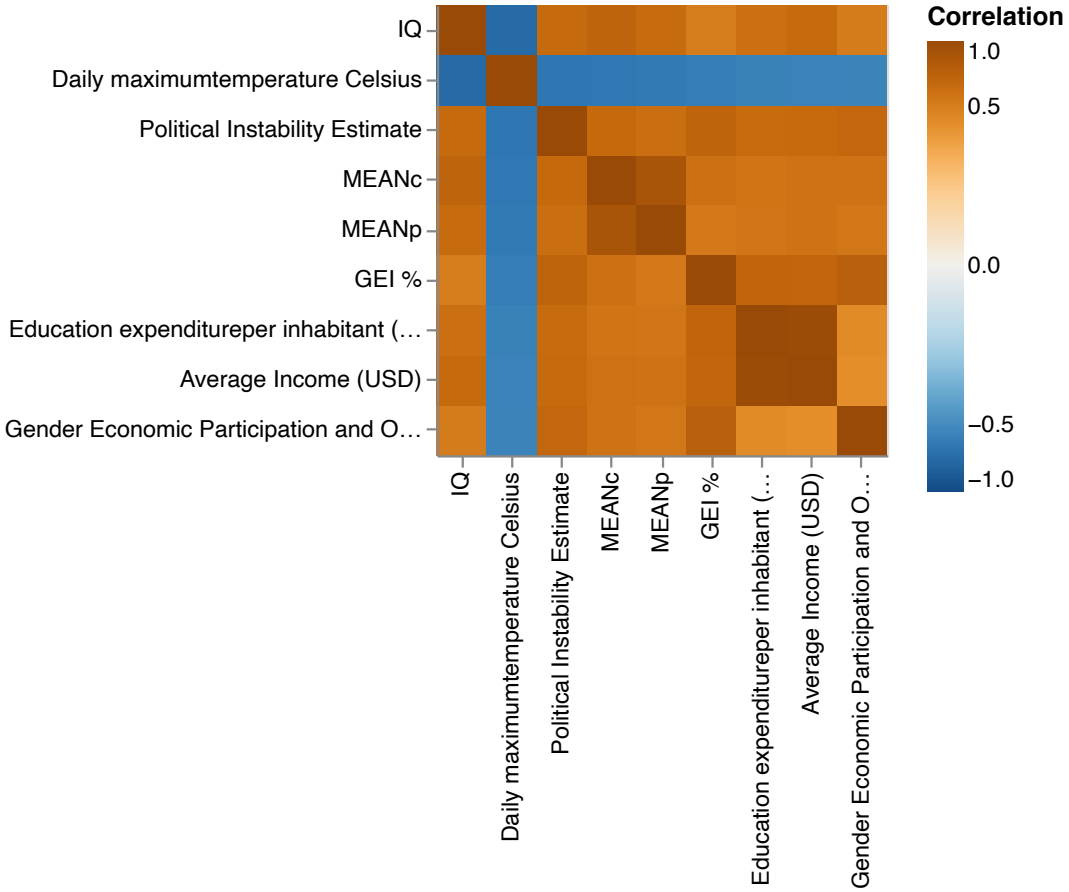


The figure farthest left is a histogram showing the average child education level by different regions and the center histogram shows the same information against sub-regions more specifically. The histogram farthest to the right shows average child education level by parent education level, distinguished by color for each established region. As we can see, regions of high income economies have higher average children education levels opposed to regions like Africa and South Asia, known for being less developed countries, have very low averages for child education level. The larger proportion of the data also comes from regions with higher economies which may be associated with the amount of accessible amenities within these regions. Since we have

found that Africa and South Asia have lower average education levels of children, it only makes sense that these regions represent smaller portions of our data. However, it is intriguing to notice a majority of parental income in South Asia are middle class income as this does not result in the a greater educational attainment for children more closely resembling regions like 'East Asia and Pacific'. This supports the conclusion that accessibility within a country is vital for educational growth.

Correlation Matrix

Principal components analysis was performed on the normalized GDIM data and identified three variables of educational mobility. These measures are weighted averages of each variable, and together capture more than 80 percent of the total variation in mean education level of children across regions worldwide.



> **Figure 2**: Correlation Matrix: a heatmap of correlation among all quantitative variables in 'GDIM' data. Color scale shows positive correlations in orange, negative ones in blue, strong correlations in dark tones, and weak correlations in light tones.

In the heatmap, we are able to see the correlation values between variables associated with gender, politics, income, and geographical influence. We did not find one significant factor that predominantly shows a larger than normal influence on educational attainment level, rather there are multiple correlated variables.The most evident

difference is the very strong negative correlation between ‘Daily maximum temperature - Celsius’ and every other variable in the correlation matrix.

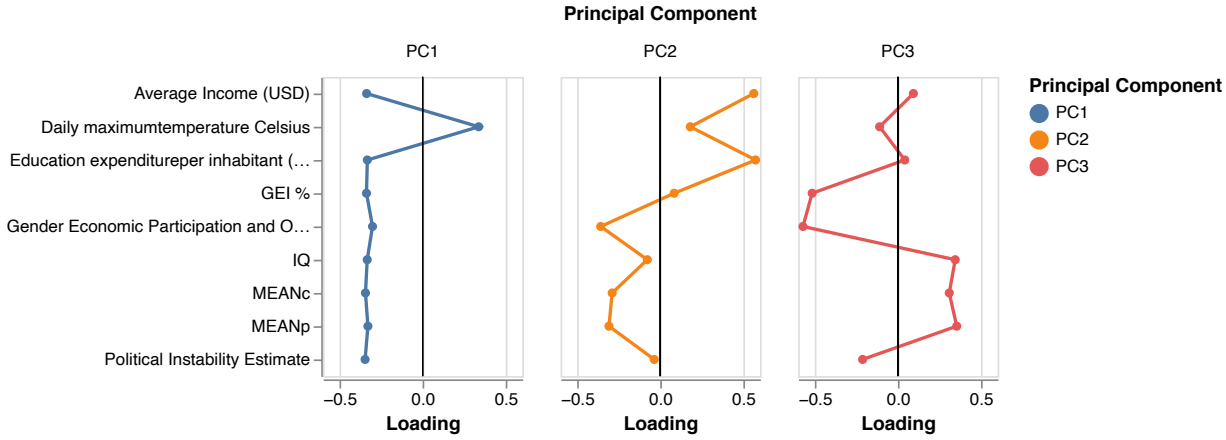
Measures of Composition

Principal components analysis was performed on the normalized GDIM data and identified three variables of educational mobility. These measures are weighted averages of each variable, and together capture more than 84 percent of the total variation in mean education level of children across regions worldwide.

> **Table 3**: Principle Component Table and Loadings.

	PC1	PC2	PC3	Variable
1	-0.344636	-0.266317	0.339438	MEANc
2	-0.331154	-0.263275	0.391683	MEANp
3	-0.335637	0.052723	-0.522737	GEI %
4	-0.299360	-0.418128	-0.533057	Gender Economic Participation and Opportunity
5	-0.345115	-0.069630	-0.220119	Political Instability Estimate

> **Figure 4**: Linear Regression Analysis.



- PC1 has **Daily temperature Celsius** as the largest positive loading with all other variables are a negative loading component. This means an increase in daily temperature (Celsius) results in a decrease of all other factors showing a negative relationship to temperature and educational growth. In addition, the data provided from PC1 also coincides with what was discovered in our correlation matrix, that daily temperature is negatively correlated to all factors.
- PC2 has **Average Income** and **Education expenditure\per inhabitant** as the largest positive loading showing these two values are correlated and an increase in these values decreases our overall **Gender Economic Participation and Opportunity Equality %**. An increase in average income or education expediture increases the value of principle component 2.

- PC3 shows **IQ**, **MEANc**, and **MEANp** as the largest positive loadings with **Gender Economic Participation and Opportunity Equality %** AND **GEI%** as the negative loadings. The same can be said if vice versa occured. If GEI% increased, this would result in a decrease in the value of PC3.

Predicting Education via linear regression

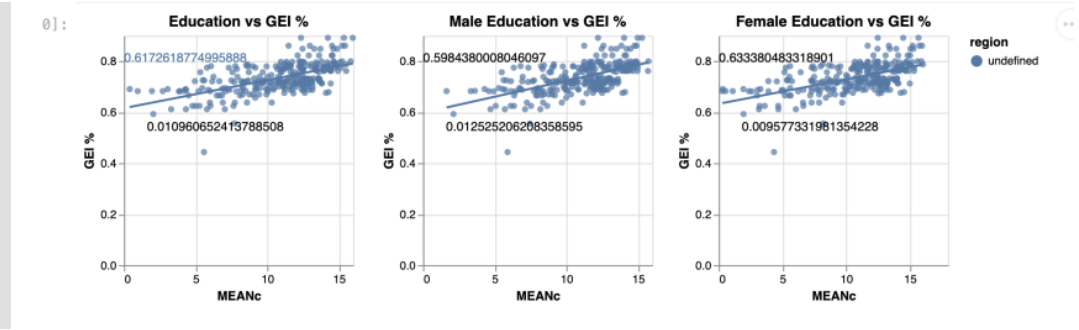
Lastly, a multiple linear regression model was performed to approximate the association the total years spent in education for children and other external factors such as average parental education, country's of residence: education expenditure, average income, gender equality, political stability, and mean temperature.

In this model all of the variables are quantitative and we have encoded all of the main variables into the regression. Table 3 shows the results of the model linear regression.

> **Table 4**: a parameter estimate table of the fitted linear regression model. the coefficient estimates and coefficient standard errors for the intercept and each variable are shown in the two columns, with rows indexed by parameter name. The estimate for the error variance parameter is in the last row.

	estimate	standard error
intercept	-8.721432	2.430557
MEANp	0.671677	0.035524
GEI%	10.179138	2.773549
Gender Economic Paraticipation & Opportunity Equality %	-1.664455	1.414770
Political Instability Esitamte	0.261234	0.171399
IQ	0.087152	0.014620
Average Income (USD)	-0.000027	0.000026
Eduncation expenditure\nper inhabitant (USD)	0.000126	0.000403
Daily maximum\ntemperature Celsius	0.036698	0.018962\
erro variance	1.803180	NaN

> **Figure 4**: This provides a regression of the variable education against GEI for all sexes with seperate regressions for male and female as well to show the difference in the Y-intercept and rate of change. There seems to be an increase in male education with increase in GEI %



In Table 4 the regression shows that a 1 unit increase in parental education results in a 0.671677 increase in child education. The standard error is .03 which leads us to conclude that it is a decent estimate as zero is not within a SDE. The estimates for all variables except the expenditure variable are all not within SDE of zero thus viable estimates, suggesting that the expenditure variable may not be significant. The GEI estimate is 10.179138 but since each increase in GEI is .01 up to 1 max, it has less of an impact then an initial brief observation of the model suggests. IQ and political instability seem to also have big impacts with each unit increase. An R^2 value of 0.8264603491332505 suggests a high amount of variance in the education can be explained by these variables.

Discussion

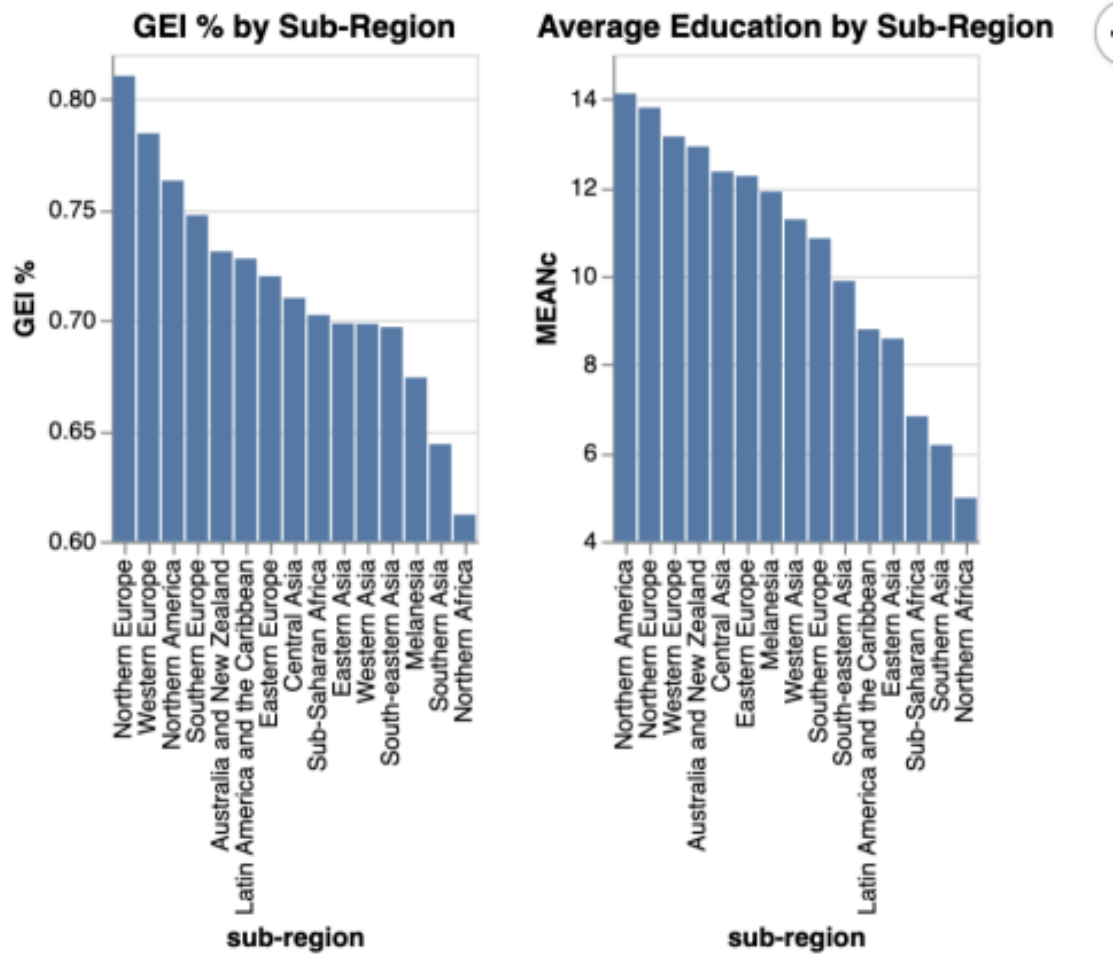
This project analyzed the correlation between education numbers and a few social, political, and economical factors. The analysis focused on different factors that influence average education of a countries residents. The analysis showed the main correlations between education levels and daily maximum temperature of your country, average income and education exepnditure per inhabitant, and also Gender Equality. These seemed to be some of the major contributors to the success of a countries education levels. Further, linear regression model quantified the relationship between these variables and the education attainment.

As seen in Figure 1, we were able to see the breakdown of child educaution levels by region and sub-region to better understand where education levels may be more prevalent than others before doing any type of anyalysis. After running a correlation matrix to find patterns in our data as well as run a PCA analysis we found that a few specific factors produced a majority of the variation in our data and untimately had the greatest influence on the overall education prospect of a given resident.

For instance, there seemed to be a unusual negative correlation between the mean years educated in a country and the average daily maximum temperature. Further research and data needs to be compiled in order to understand why this is the case. Mean temperature also seems to have a correlation with average income and political instability thus more reasearch would need to be done to understand this issue. We

found other variables such as average income of a resident and the amount of educational expediture spent by such resident go hand in hand and influencing education levels. Factors like IQ and MEANp (parent schooling) also work hand in hand with educational growth.

[11]:



Another surprising find is that there seems to be an increase in likelihood of higher education for men when gender equality is higher (Figure 4). This is shown by an increase in the slope of the regression line when education is plotted against GEI % for all, males and females. Whilst a higher GEI % increases the overall effect of higher GEI % for men, it does the exact opposite for women. This is inline with current studies stating that as Gender equality index increases, the biological differences between men and women tend to manifest themselves more so, possibly resulting in different educational attainment goals for the different sexes.