Alex Scarlatos
CSE 564
Dr. Klaus Mueller
5/12/17

Effects of Education on the Well-Being of Nations

Proposal:

Motivation

Improving education has always been a major goal of philanthropists and world organizations when attempting to aid poor nations. In first world countries, we believe that a good education is key to our way of life and high standard of living. But it is important to examine exactly what role education has in the well-being of a country. Certain aspects of a country could improve due to providing education, while others could be dependent on other factors. If organizations are to determine where education aid should be going, they need to know about strong correlations between aspects of a country's well-being and the education of its people. We can discover which countries should be targeted for education aid by examining global and localized trends, and comparing them to other cases across the world.

Data Source

I will be using the World Bank as the source for my data. They have a database of a huge variety of attributes for countries around the world over a range of years. Their data is formatted so that each country is a row, each year is a column, and each cell is the value of the variable in question. I plan to create a new dataset using these preexisting ones by making country and year into a key pair, and making a column for each variable I want to examine.

I can prepare several datasets using this method in order to compare correlations between different areas. The World Bank has groups of variables such as health, aid effectiveness and poverty. Each of these groups of variables could be made into its own dataset, and combinations of groups, for example education and poverty, could be useful when trying to focus in on specific correlations. These will be very high dimensional datasets (some of the groups have more than 50 attributes) so I will need to use judgement to make initial reductions, and then PCA or other methods to further reduce the dimensionality.

Analysis Plan

After finding the variables with the strongest PCA loadings in each dataset, I will combine them into a single large dataset with the most important variables from each relevant group. Then I'll compare the correlations between these variables and visualize them with MDS. Since the datasets provide years, I will watch the correlations evolve over time, and the way that a variable's correlation with education changes will be revealing. It could indicate a false positive if the correlation moves away from education over time but is more likely to be relevant if it stays close over time.

I plan to do clustering on the countries to find out which ones have similar attributes. It will be useful to compare trends between countries within clusters. For example, maybe one group of countries has poor infrastructure, and within that group education benefits are different from countries with better infrastructure. Within these clusters I'll try to find the strongest correlated variables to education and create 2D visualizations across those variables. I will also identify attributes that set clusters apart from each other. If the clusters show different correlations with education, then I will be able to identify attributes that change the way that education affects or is affected by a country's status.

Some deviations may be hard to explain, so I will identify outlying countries and attempt to explain their peculiarities with data that is not found in the dataset. These may have to do with politics or internal conflict.

By the end of these analyses, I will have identified variables that are strongly correlated to education in countries under different conditions. I expect some of these correlations to vary across different regions of the world, and others to be more global.

Plan for Final Presentation

After discovering which attributes and regions are worth examining, I plan to present evidence of these trends along with future predictions and analysis.

The evidence will comprise a variety of visualizations, including 2D attribute charts, 2D artificial dimension charts (like MDS) and higher dimension charts (like parallel coordinates). I will use analysis techniques, like traveling salesman, to find the best way to present these results.

I will also use a world map to show the most relevant attributes and relations. It will be useful to display how certain attributes correlate with education differently across the world. For example, if health, poverty and infrastructure vary with their education correlations across different groups of countries, I will assign those attributes colors and fill in the map accordingly.

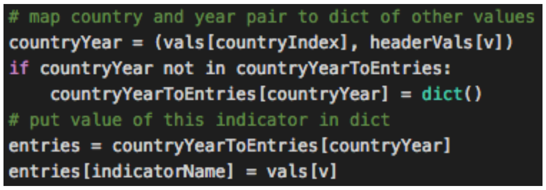
Based on these results, I will extrapolate the strongest trends to make predictions about the future conditions of education and well-being for countries that are going through clear changes. I will also use the results to determine which countries could benefit from education by comparing them to similar countries that have improved from it. This could be useful for organizations in determining which countries would benefit the most from education aid.

Preliminary Report:

Approach and Process

I initially had to find a set of variables to compare from the World Bank's vast selection. I chose several education variables that I thought would be good indicators, like literacy rate and primary completion rate, along with many other variables that I thought could be affected by or affect the education variables. They ranged from poverty headcount to high-tech exports, and I ended up with an initial list of about 30 variables.

The World Bank formats their data such that each variable gets a whole table, where each country gets a row, each year gets a column, and the value of the variable in question goes in the cell. I wrote a Python script to iterate over many of these tables and compile them into one. It uses a (country, year) tuple as a key for a dictionary, whose value is another dictionary that maps variable name to value.

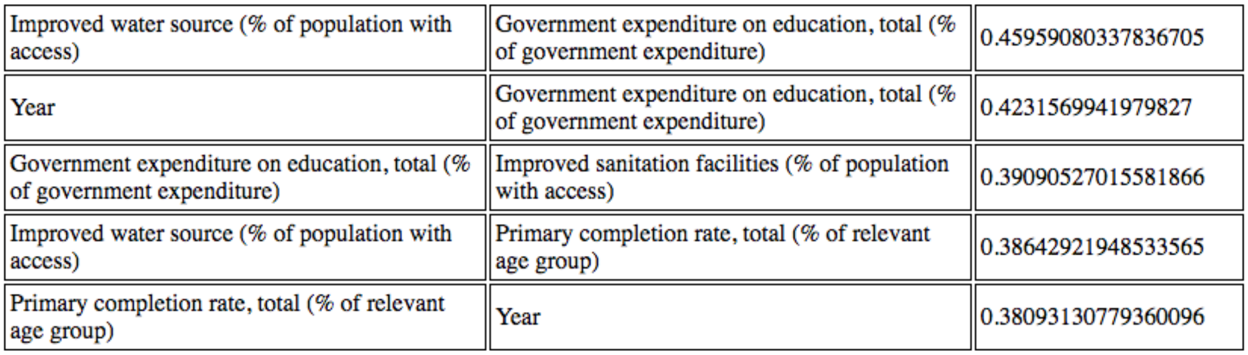


```
# map country and year pair to dict of other values
countryYear = (vals[countryIndex], headerVals[v])
if countryYear not in countryYearToEntries:
    countryYearToEntries[countryYear] = dict()
# put value of this indicator in dict
entries = countryYearToEntries[countryYear]
entries[indicatorName] = vals[v]
```

Figure 1 - Called for every column (v) of every row (vals) of every source file

To analyze the data, I used the app that I built for Mini Project 2, which uses a Python Flask server for data operations and d3js and HTML for data visualization.

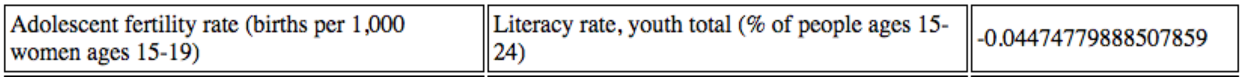
To get an idea of which variables were most relevant, I needed to find the highest correlations of education variables to the other variables. I generated a correlation matrix, and filtered it to only show correlations with education variables, and sorted it by magnitude.



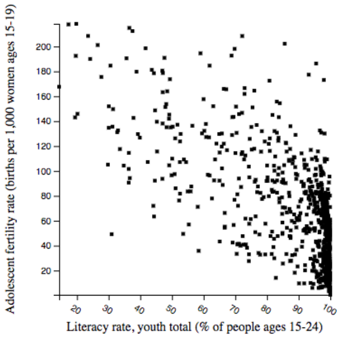
Improved water source (% of population with access)	Government expenditure on education, total (% of government expenditure)	0.45959080337836705
Year	Government expenditure on education, total (% of government expenditure)	0.4231569941979827
Government expenditure on education, total (% of government expenditure)	Improved sanitation facilities (% of population with access)	0.39090527015581866
Improved water source (% of population with access)	Primary completion rate, total (% of relevant age group)	0.38642921948533565
Primary completion rate, total (% of relevant age group)	Year	0.38093130779360096

Figure 2 - The first 5 non-inter-education-variable correlations

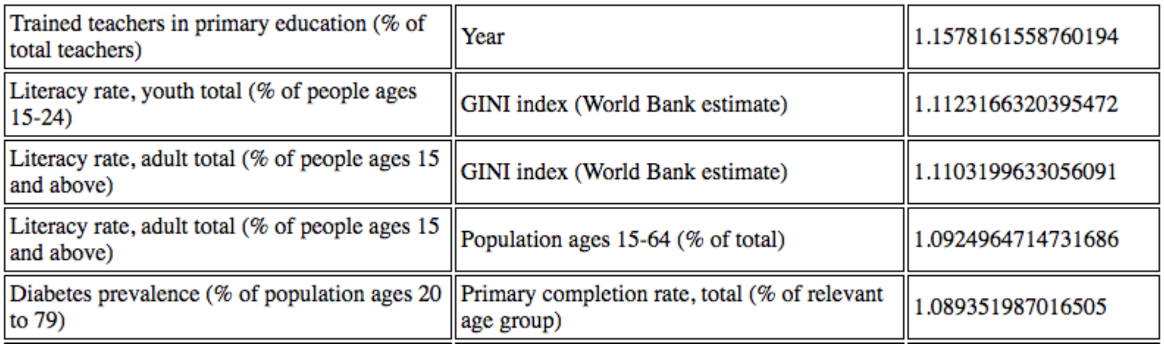
But I noticed that even small correlations showed clear relationships.



Adolescent fertility rate (births per 1,000 women ages 15-19)	Literacy rate, youth total (% of people ages 15-24)	-0.04474779888507859



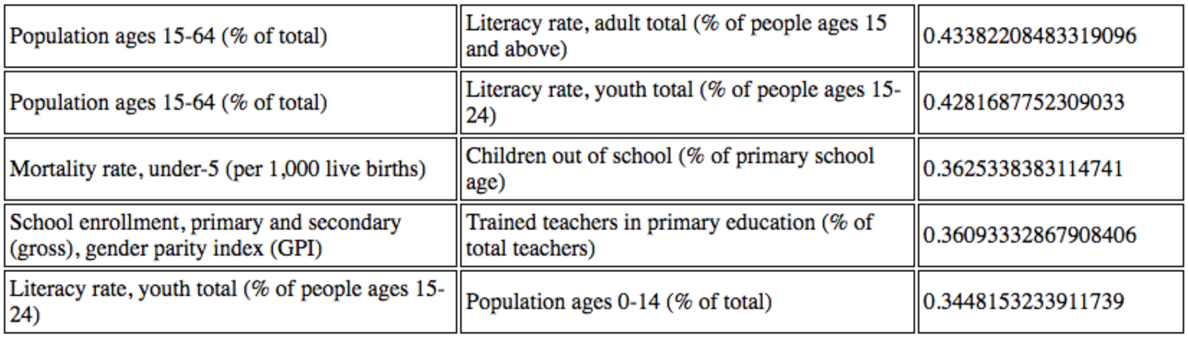
To account for this, I decided it would be best to examine correlations on more local scales, like within a small group of countries or during a small group of years. This could be done manually or with clustering, so I tried both.



Trained teachers in primary education (% of total teachers)	Year	1.1578161558760194
Literacy rate, youth total (% of people ages 15-24)	GINI index (World Bank estimate)	1.1123166320395472
Literacy rate, adult total (% of people ages 15 and above)	GINI index (World Bank estimate)	1.1103199633056091
Literacy rate, adult total (% of people ages 15 and above)	Population ages 15-64 (% of total)	1.0924964714731686
Diabetes prevalence (% of population ages 20 to 79)	Primary completion rate, total (% of relevant age group)	1.089351987016505

Figure 3 - Top correlations after stratified sampling based on k-means clustering

For more manual analysis, I made it so the user could filter the data on year, country and region, since I expected to see correlations change over time and be different around the world.



Population ages 15-64 (% of total)	Literacy rate, adult total (% of people ages 15 and above)	0.43382208483319096
Population ages 15-64 (% of total)	Literacy rate, youth total (% of people ages 15-24)	0.4281687752309033
Mortality rate, under-5 (per 1,000 live births)	Children out of school (% of primary school age)	0.3625338383114741
School enrollment, primary and secondary (gross), gender parity index (GPI)	Trained teachers in primary education (% of total teachers)	0.36093332867908406
Literacy rate, youth total (% of people ages 15-24)	Population ages 0-14 (% of total)	0.3448153233911739

Figure 4 - Correlations in similar range in 2015 only

The clear differences indicate that I will have to exhaustively examine different groupings of the data to find clear patterns. I will need better visualizations to do so, which is something I will implement in the next part.

Another factor that could be leading to unstable correlations is the fact that the World Bank datasets are not complete, and are missing many values for different (country, year) pairs. I needed to account for this in plotting the data and performing operations on it.

I noticed with some of the very high correlations, only a handful of rows existed with entries for the two variables. To avoid giving these uncertain variables too much weight in the final analysis, I decided to assign each correlation a “reliability” status, proportionate to the number of available points relative to the total number of points. This status will be used later to determine which correlations are worth examining.

Plan for Final Report

Before continuing, I will first have to examine the correlations closely to reduce the number of variables to make visualization and data manipulation quicker and to focus in on certain variables.

I plan on taking the correlations I found and visualizing them in more effective ways. I also plan to provide a better system for user interaction, so that one can see patterns that go beyond the individual correlations.

I want to generate biplots, based on PCA analysis, showing data points and attribute vectors mapped onto the top 2 PCA vectors. The data points could be color-coded on different classifications, like region, or clusters, so users can see patterns.

I also want to generate an MDS chart on the attributes, so users can see all the correlations at once.

Finally, I want to show a world map, where each country can be color-coded based on attribute values and correlation values.

All of the above visualizations will be interactive in that users will be able to select countries, regions, years and other variables to filter on and see changes in real time. They will also be able to hover over items to see relevant attributes.

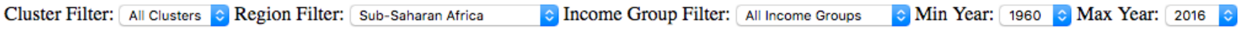
Using the visualizations, I expect to find patterns among regions of the world, groups of countries (identified by clusters or predefined World Bank groups), and in changes over time. I will gather the best group of patterns I find, save the visualizations that presented them, and use them to come up with a hypothesis as to which aspects are most strongly affected by or affect education.

Final Report:

Software Overview

The software allows users to explore different areas in the data to find interesting correlations. There are two pages: one for the main visualizations and filtering, and one for data sampling and viewing data points by attribute.

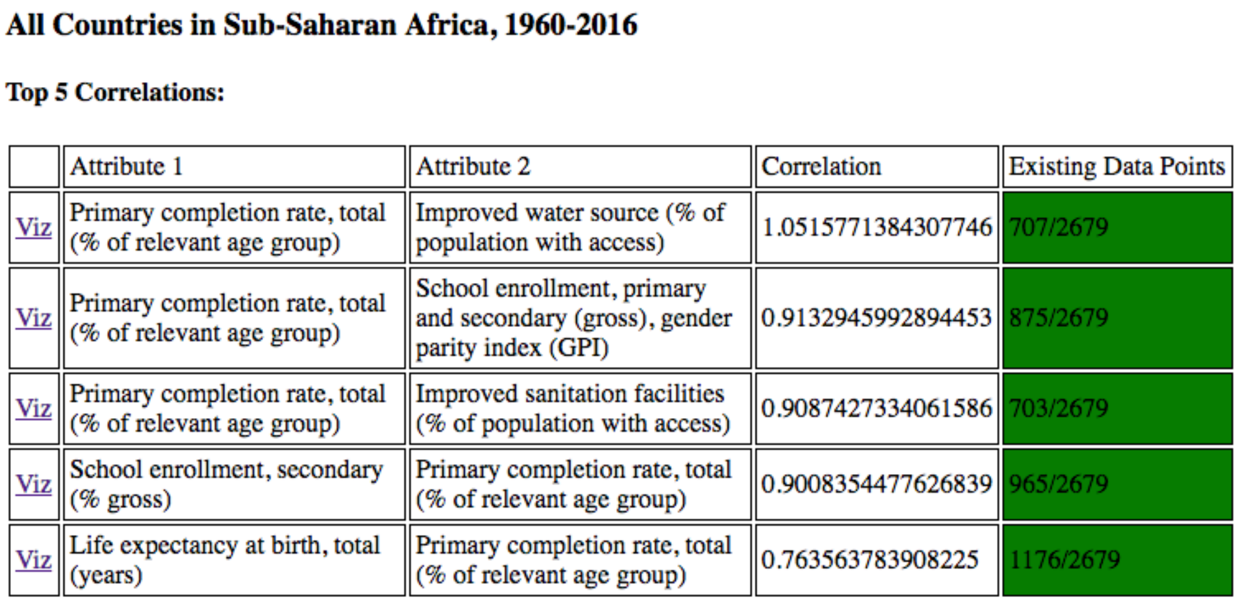
The first page allows users to filter by cluster, region, income group and year.



Cluster Filter: All Clusters Region Filter: Sub-Saharan Africa Income Group Filter: All Income Groups Min Year: 1960 Max Year: 2016

The clusters are found by k-means clustering, based on Euclidean distance across all attributes. The default max-k is set at 2, but can be increased for more interesting results. Region and income group are predefined in the World Bank dataset.

Below the settings, you see an overview of the top correlations.

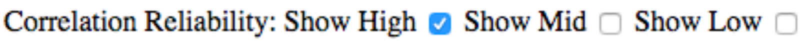


All Countries in Sub-Saharan Africa, 1960-2016

Top 5 Correlations:

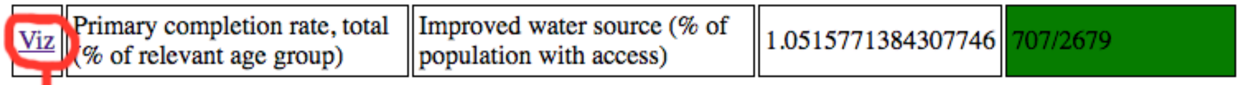
	Attribute 1	Attribute 2	Correlation	Existing Data Points
Viz	Primary completion rate, total (% of relevant age group)	Improved water source (% of population with access)	1.0515771384307746	707/2679
Viz	Primary completion rate, total (% of relevant age group)	School enrollment, primary and secondary (gross), gender parity index (GPI)	0.9132945992894453	875/2679
Viz	Primary completion rate, total (% of relevant age group)	Improved sanitation facilities (% of population with access)	0.9087427334061586	703/2679
Viz	School enrollment, secondary (% gross)	Primary completion rate, total (% of relevant age group)	0.9008354477626839	965/2679
Viz	Life expectancy at birth, total (years)	Primary completion rate, total (% of relevant age group)	0.763563783908225	1176/2679

The title is generated based on your filter settings. The correlations are the top 5 that include educational variables. They are ranked based on the existing number of data points, since some entries are missing in the dataset. The ranking is referred to as “reliability”, and you can filter out correlations of different reliabilities. High has > 25% of possible entries, mid has > 10% and low has < 10%.

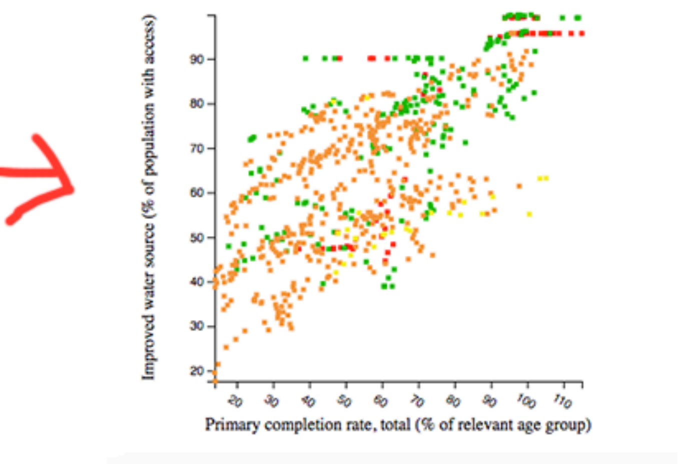


Correlation Reliability: Show High ☑ Show Mid ☐ Show Low ☐

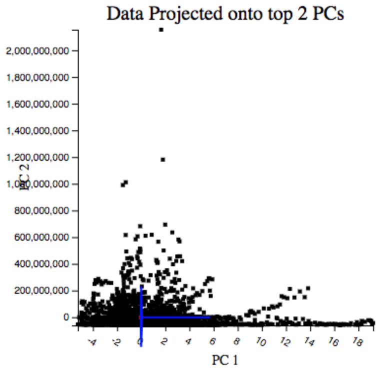
You can click on the “Viz” link to view a graph of the data points in the selection on the two attributes of that row.

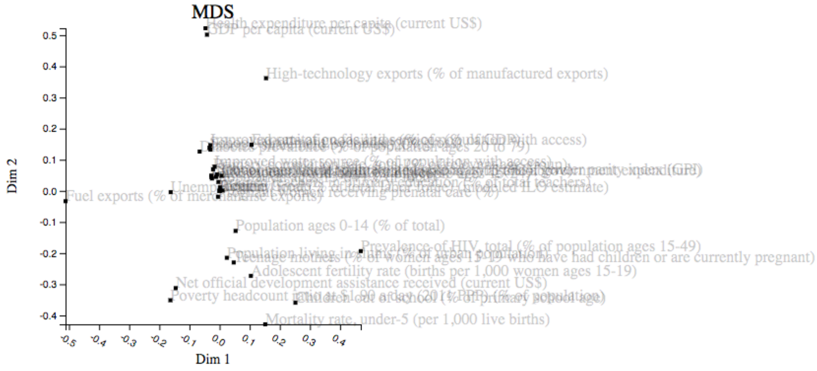


Viz	Primary completion rate, total (% of relevant age group)	Improved water source (% of population with access)	1.0515771384307746	707/2679



Below this overview section are two graphs. The first is a biplot, showing all data points mapped onto the top 2 principal components of the dataset. Attribute vectors are also mapped onto the graph. You can mouse over data points to see the (country, year) pair represented, and you can mouse over attribute vectors to see their name as well. The second graph shows the correlations of all variables in the dataset. Higher correlations are put closer together, and the graph was generated using MDS.





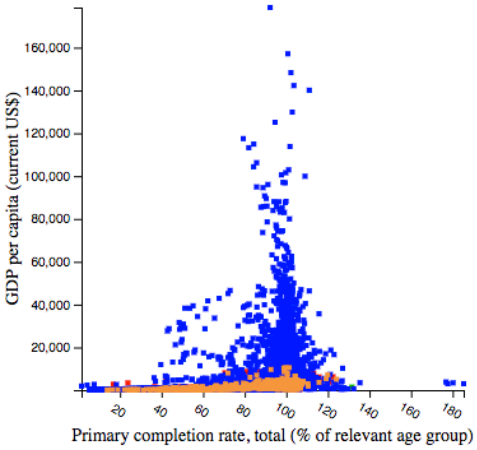
These graphs are very useful for finding outliers. On the PC graph, you can clearly see that several data points stray from the main cluster. By mousing over the points, it is found that the points up high are Nigeria and DRC, the points at the far right are Mauritius and Seychelles, and the small cluster above the tail on the right consists of South Africa. This indicates that these countries might be worth further examination, or that they could be disregarded when examining attribute pair graphs.

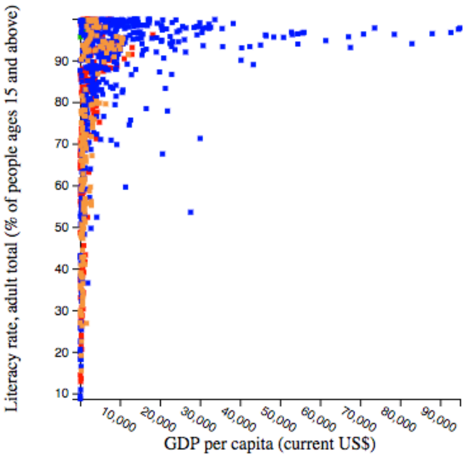
Since there are many education variables, they tend to be clustered together in the center of the MDS graph. Because of this, the graph is useful for seeing which attributes are outliers (not highly correlated to any education attributes). These can generally be disregarded and considered to not be heavily affecting education in the selection.

Finally, below these two graphs is a table of all correlations with education variables, similar to the one seen above but not restricted to just 5.

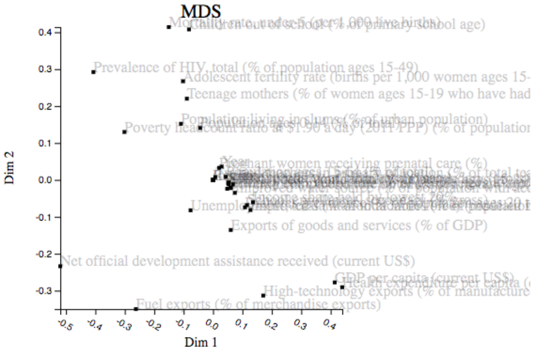
Initial Findings and Reductions

I observed the correlations of the full dataset to initially remove some useless attributes. I noticed that inflation as a GDP deflator, and ODA (official development assistance) per capita received, always produced correlations < .1 with education attributes. Also, surprisingly, GDP per capita tended to produce lower correlations as well, but when I observed the graphs I saw that this was specific to clusters.





So after removing the lowest ranked attributes and removing ones with sparse data, I examined by cluster. I used MDS to quickly see which attributes were outliers and therefore could be thrown away.



Outlier Attributes:
All - Net official development assistance received, Fuel exports, Mortality rate under 5, Children out of school, Prevalence of HIV
Cluster 1 - Fuel exports, Prevalence of HIV, Poverty headcount ratio at $1.90 a day
Cluster 2 - Fuel exports, Health expenditure per capita, GDP per capita, Children out of school, Prevalence of HIV
Cluster 3 - Prevalence of HIV, Mortality rate

The only consistent outlier was prevalence of HIV, which I found surprising so I decided to keep it since I thought it provided insight. I also noticed that children out of school tended to form its own cluster, which I also found interesting.

Final Analysis
I decided to examine correlations by region to get a feel for how they differed and for general trends.

Most relevant variables (ones appearing in top 5 correlations) for:
Sub-Saharan Africa – Primary completion rate, Improved water source, Improved sanitation facilities, School enrollment GPI, Secondary school enrollment, Life expectancy
South Asia – Primary completion rate, Improved water source, School enrollment GPI, Government expenditure on education, Life expectancy, Population ages 15-64
Europe & Central Asia – Life expectancy, School enrollment GPI, Population ages 15-64, Primary completion rate, Improved water source
Middle East & North Africa – Improved water source, Secondary enrollment GPI, Improved sanitation facilities, Primary completion rate
East Asia & Pacific – Primary completion rate, Population ages 15-64, Secondary enrollment GPI, Life expectancy, Secondary school enrollment
Latin America & Caribbean – Secondary enrollment GPI, Life expectancy, Primary completion rate, Population ages 15-64
North America – Life expectancy, Government expenditure on education, Secondary enrollment GPI, Population ages 15-64, Secondary school enrollment, High-tech exports

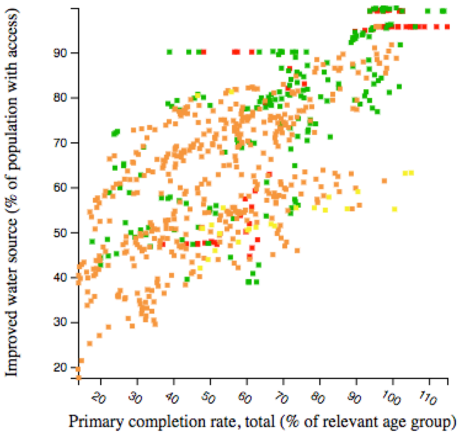
There were some variables that clearly made many appearances here. Primary completion rate (the % of students that complete primary school) and secondary enrollment GPI (the ratio of women to men in secondary school) seemed to be the strongest educational variables.

Some variables were more prominent in certain areas, like improved water source and sanitation facilities in Eurasia and Africa. However, life expectancy and population ages 15-64 were consistently highly correlated with educational variables across the globe.

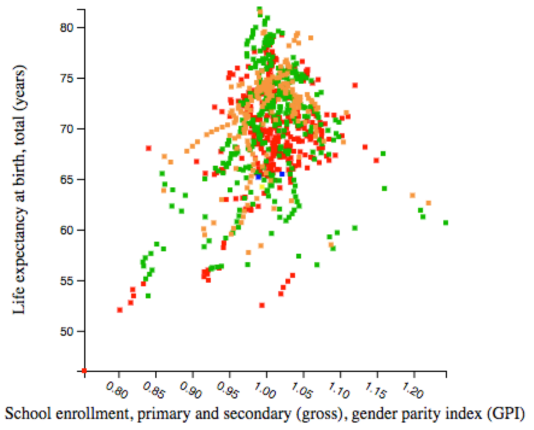
Additionally, North America seemed to be an outlier in all regards, possibly because there were many less data points than in other regions, and because of the region's relatively high wealth.

I decided to map out some of the highest correlations that I found interesting.

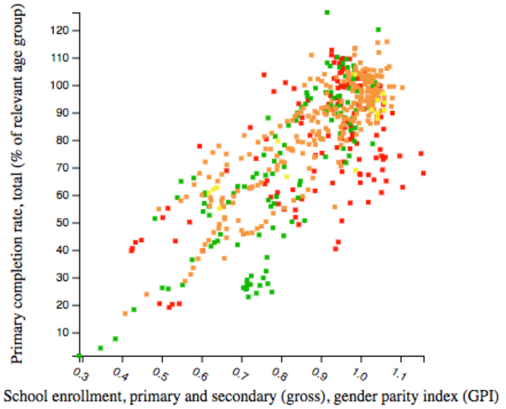
In Sub-Saharan Africa:



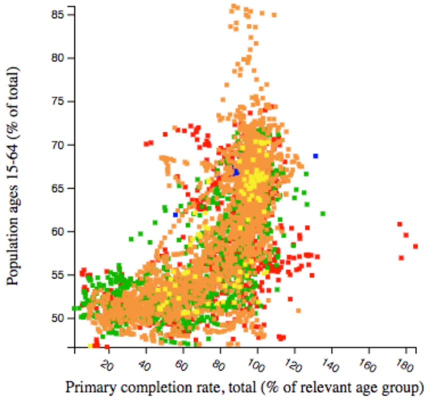
In Latin America & Caribbean



In Middle East & North Africa:



In World:



Research Conclusion

In under-developed regions of the world, education increases greatly with the implementation of necessary infrastructure that provides water and sanitation.

Throughout the world, life expectancy rises greatly with education. Additionally, a barrier between some countries and well-being is gender equality in education, as it seems that countries with unequal gender enrollment rates are restricted to low life expectancies and low completion rates.

Data Resource

“Indicators.” *The World Bank*, http://data.worldbank.org/indicator.