Education and International Relations

A Data Analysis Project

Panagiotis Podiotis

Department of International and European Studies

University of Macedonia

Thessaloniki, Greece

May 30, 2020

**Prepared for the course "Special Topics of International and European Educational Policies.*

Supervisor: Prof. Boutsiouki Sofia.

***Code and data can be found in: https://github.com/Podipan/cia_world_factbook*

Contents

Introduction 2
This Paper 3
The Data 4
Correlation Analysis 8
Clustering 18
Conclusion & Ideas for further research. 20
References 21

Tables & Figures

Figure 1 - PISA Score and School Life Expectancy with best fit line. 4
Figure 2 – UN Education Index and School Life Expectancy with best fit line. 5
Figure 3 - Correlation Coefficients for CIA World Factbook's School Life Expectancy Data. 5
Figure 4 - PISA Score and GDP Education expenditure with best fit line. 6
Figure 5 - UN Education Index and GDP Education expenditure with best fit line. 6
Figure 6 - Elbow method for optimal number of clusters. 18
Figure 7 – Silhouette method (index) for optimal number of clusters. 18

Table 1 - School Life Expectancy as a measure of education quality, correlations 5
Table 2 - Education Expenditures as a measure of education quality, correlations 7
Table 3 - CIA World Factbook Education correlations. 8
Table 4 - Clustering Metrics. 19
Table 5 - Clustering Outcomes. 19

Introduction

Education has always been an indispensable element of human society and as a consequence, an item of study for International Relations as well. The development of mankind unravels in parallel with the increase of education's value. Education, now more than ever, seems to be the answer to major contemporary challenges like hybrid warfare, technological antagonism and globalization. It is within schools and universities where analytical thinking, technological literation and research occurs. It is education which builds critical thinking and identity as a defense against disinformation.

For many years, governments have been interacting with one another, nowadays, governments have the ability to address one another's citizens directly attempting to promote their narratives. Education is probably the most robust mean of communication between a government and its own citizens.

It is the situation described above where this paper comes into existence. On the intersection of International Relations, Data Science and Education. It could thus be characterized as a *"Computational International Relations"* (Unver, 2018) paper, a contemporary approach to modern problems.

This Paper

The present paper can be considered a Data Analysis project. Education-related data within the CIA World Factbook (hereinafter Factbook) (CIA, 2019, p. About) will be explored. The Factbook[1] has been chosen as the source of data for this project for the following reasons:

1. It provides a big volume of data across a variety of sectors[2].
2. High confidence can be placed on the data since the CIA is one of the most recognized and prestigious US governmental agencies specializing on intelligence.
3. Data are open-source and available for download.
4. The author has worked extensively with the Factbook by having converted it in dataset format (Podiotis, 2020)[3] and by having conducted various ML sub-projects with it.
5. There is a general lack of extensive datasets which cover International Relation topics in such a wholistic manner.

Data Science provides with almost endless research approaches. In an effort to limit the scope of the paper and increase the quality of research, it was decided that two tasks will be performed:

1. Correlation exploration.
2. Clustering.

These two approaches cover the first steps of data analysis. The author hopes to lay the ground for future research and hypothesis with the findings of this paper.

All of the above tasks will be carried out with the help of Python 3.7 & 3.8 (Python Foundation, 2001) programming language, within the JetBrains PyCharm Community Edition 2019.2.2 x64 (JetBrains, 2000) Integrated Development Environment.

[1] The latest available version (2018 version, uploaded on Jan 04, 2019 01:51 PM – downloaded 15/05/2020) of the CIA World Factbook which was downloaded through the official web portal was used.
[2] The Factbook covers 12 thematic areas for each entry (country), namely: History, People and Society, Government, Economy, Energy, Geography, Communications, Transportation, Military, Terrorism, Transnational Issues. Each Thematic area (category) consists of numerous fields providing numeric or textual data points.
[3] Pending title and official submission.

The Data

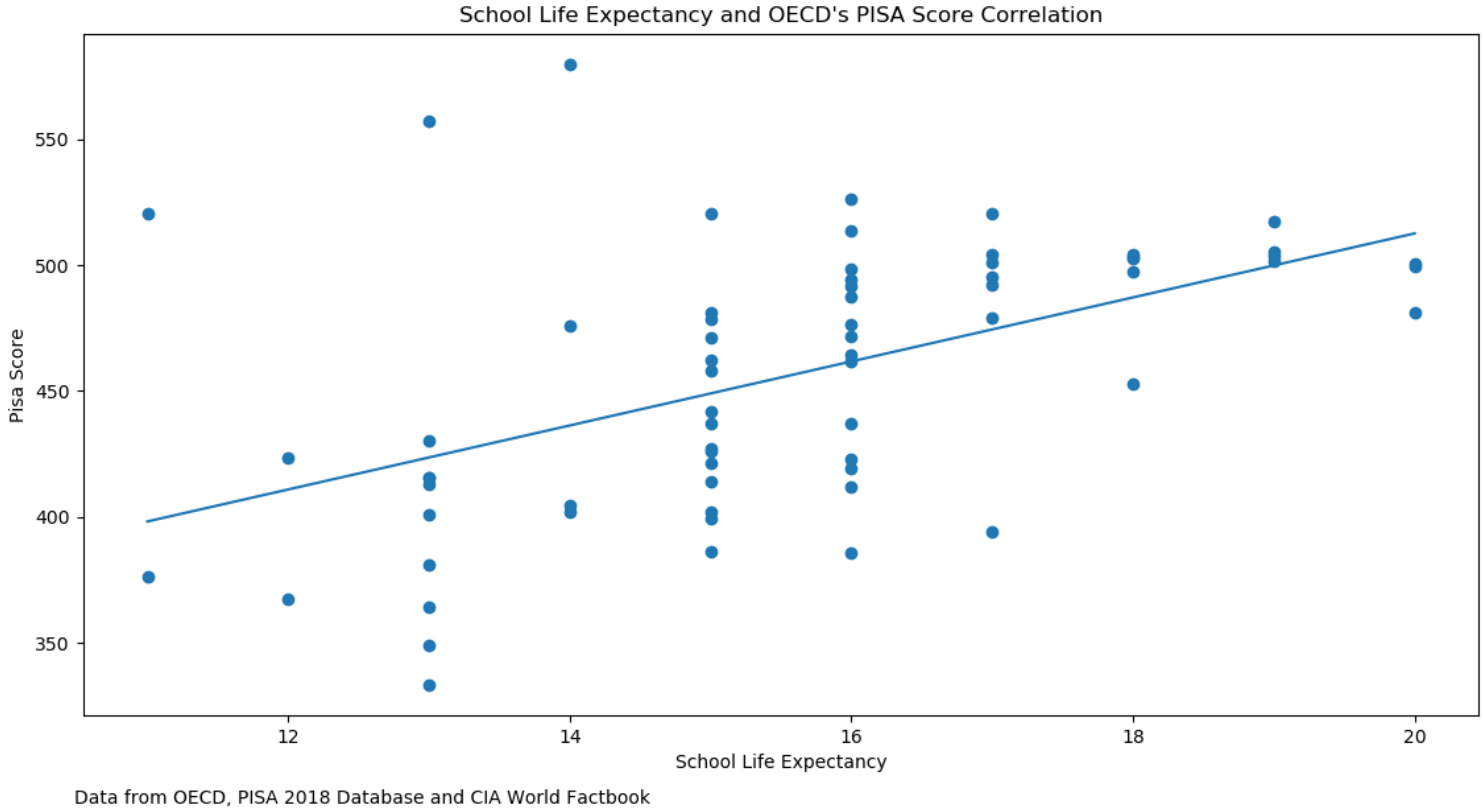
The Factbook contains both numerical and textual data. Out of all data fields, four are specifically related to education. These fields are: education-expenditures, school-life-expectancy-primary-to-tertiary-education male, school-life-expectancy-primary-to-tertiary-education female, school-life-expectancy-primary-to-tertiary-education total. The three "life-expectancy" fields originally contained missing values which were imputed by the author in the context of another project (Podiotis, 2020). The Mean Average Percentage Error of imputed values is 14%, 7% and 3% respectively.

"School life expectancy (SLE) is the total number of years of schooling (primary to tertiary) that a child can expect to receive, assuming that the probability of his or her being enrolled in school at any particular future age is equal to the current enrollment ratio at that age...SLE represents the expected number of years of schooling that will be completed, including years spent repeating one or more grades." (CIA, 2019)

Kolmogorov-Smirnov test[4] was conducted iteratively for various distributions on the values of the education-related data (columns) in order to establish their distribution. All four were found to be normal distributions. Despite that, not every other column in the dataset follows a normal distribution.

Before proceeding with this paper, it is crucial that the meaning of the School Life Expectancy data is understood. Can we generalize School Life Expectancy data to represent a country's overall education quality? The CIA notes that: "*Caution must be maintained when utilizing this indicator in international comparisons. For example, a year or grade completed in one country is not necessarily the same in terms of educational content or quality as a year or grade completed in another country".* This definition raises a warning which refers to the comparative study between countries. In order to determine whether Factbook's School Life Expectancy entry can represent the overall quality of education in a country, the School Life expectancy was studied along with two of the most recognized and accepted quality of education indicators: OECD's PISA index (OECD, 2018) and UN's Education Index (UNDP, 2018).

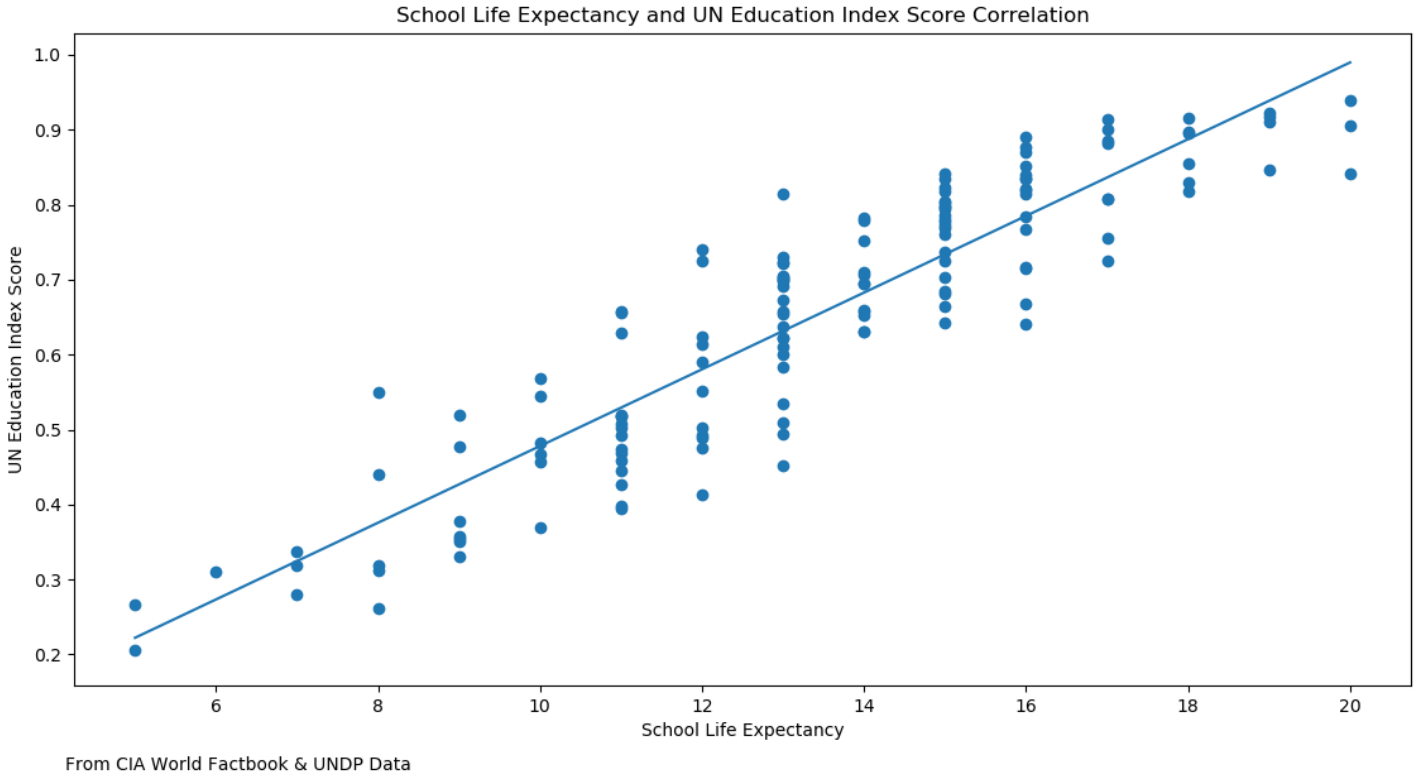
Figure 1 - PISA Score and School Life Expectancy with best fit line.



[4] Modified for goodness of fit. Implemented with Scipy (Eric, et al., 2001), function written by (Tennage, 2018).

The UN Education Index covers 145 countries and accounts socioeconomic factors as well. Thus, the UN's index is more inclusive.

Figure 2 – UN Education Index and School Life Expectancy with best fit line.



From all the data presented above, it can be assumed that the Factbook's School Life Expectancy data can represent the quality of education in a country with high accuracy.

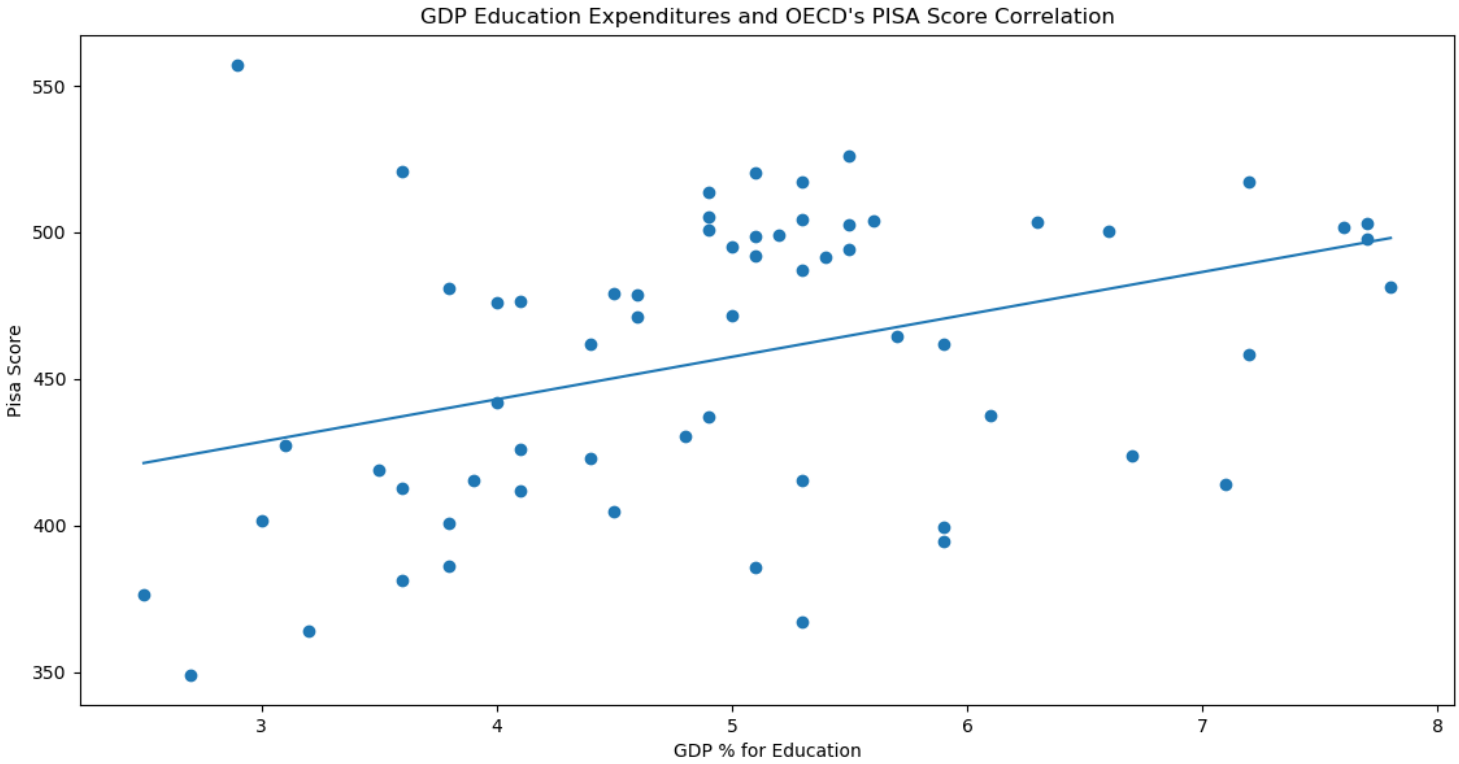
Table 1 - School Life Expectancy as a measure of education quality, correlations

	Total School Life Expectancy		
Correlation Coefficients:	Pearson's r	Spearman's r	Kendall's τ
UN's Education Index	0.90	0.90	0.76
OECD's PISA Score	0.51	0.55	0.43
			$p < 0.05$

The correlation coefficients further support the hypothesis that School Life Expectancy is a strong indicator of quality of education in a country.

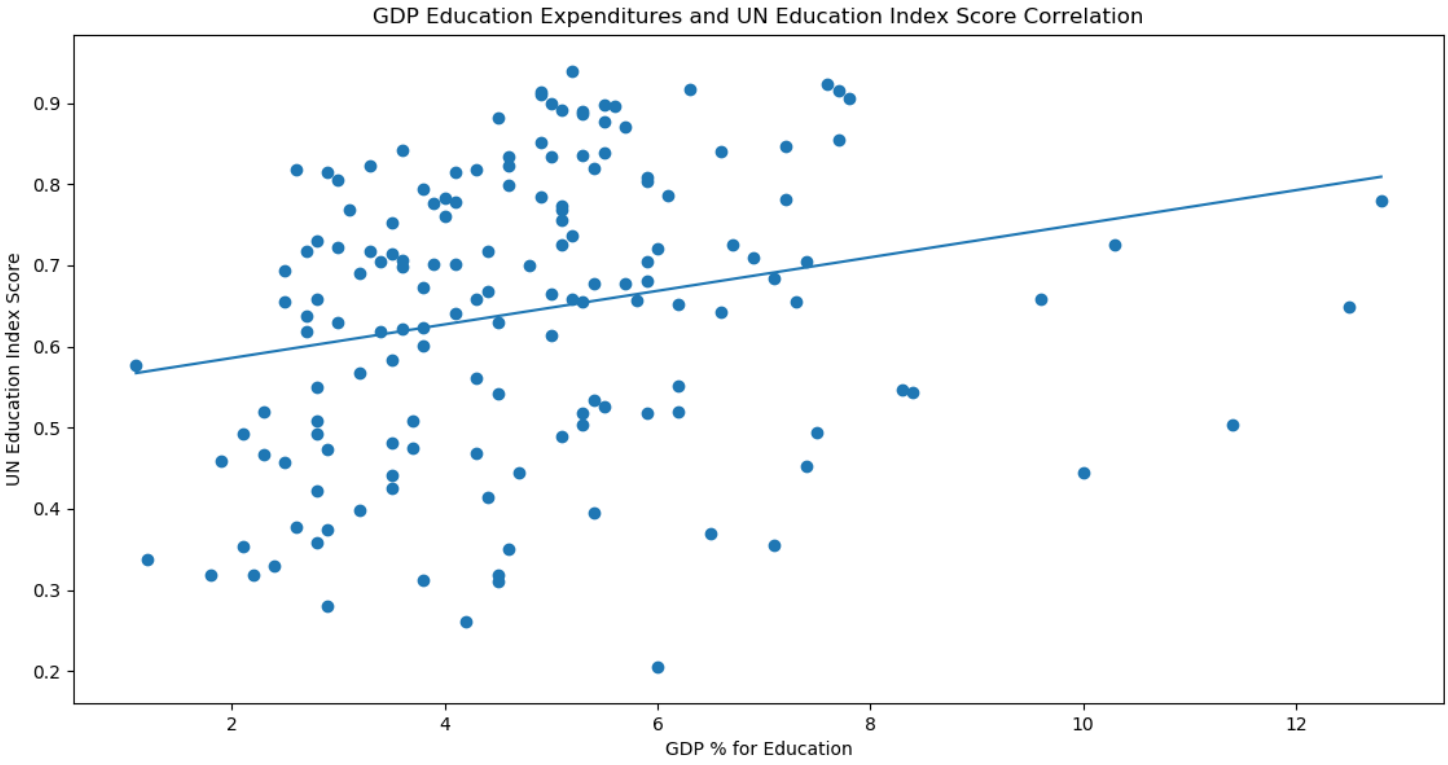
Moving on to the last education-related column of the CIA World Factbook, "education-expenditures" which provides with the GDP percentage of each country invested in education. Can it accurately represent quality of education? The same evaluation approach with the "school-life-expectancy" was used. Specifically, comparison with the PISA and UN Education Index.

Figure 4 - PISA Score and GDP Education expenditure with best fit line.



Data from OECD, PISA 2018 Database and CIA World Factbook

Figure 5 - UN Education Index and GDP Education expenditure with best fit line.



From CIA World Factbook & UNDP Data

Table 2 - Education Expenditures as a measure of education quality, correlations

	Education Expenditures (GDP %)		
Correlation Coefficients:	Pearson's r	Spearman's r	Kendall's τ
UN's Education Index	0.23	0.30	0.21
OECD's PISA Score	0.37	0.37	0.27
			$p < 0.05$

Surprisingly, education expenditures as percentage of GDP do not seem to reflect quality of education in a confident manner. This becomes evident by both the visualizations and correlation coefficients.

Correlation Analysis

It is essential for the non-technical reader to note that correlation does not mean causation. Correlation is merely a statistical observation which simplistically uncovers the existence of relation between two sets of values. It can help us increase our understanding but cannot show which variable affects the other.

In order to form high confidence results, both parametric and non-parametric methods were used. Specifically, Spearman's and Pearson's r correlation coefficients. Results were only accepted if $p<0.05$. Python implementation was carried out with Scipy (Eric, et al., 2001). As mentioned before, the dataset contains two different groups of education-related columns: Group 1 consists of three distinct columns (male, female, total) each describing "school-life-expectancy-primary-to-tertiary-education" and group 2 contains only 1 column, "education-expenditures". Both of these groups were analyzed and will be discussed in the respective order.

The primary findings of the correlation calculations between the columns of the first group and of all of Factbook's available columns yielded the below results. The correlation threshold for Group 1 was set at 0.5 (-1 to 1 scale) in order to examine strong correlations.

Can School Life Expectancy data be generalized to express overall quality of education? The CIA warns that *"Caution must be maintained when utilizing this indicator in international comparisons. For example, a year or grade completed in one country is not necessarily the same in terms of educational content or quality as a year or grade completed in another country"*. It should be noted though that in this paper the quality of education is approached holistically and not comparatively between countries. Moreover, considering the 25 countries

Table 3 - CIA World Factbook Education correlations.

	School Life Expectancy Primary to Tertiary Education *"represents the expected number of years of schooling that will be completed, including years spent repeating one or more grades"* (CIA, 2019)						
	Male		Female		Total		
Correlation Coefficients:	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Comments
major-infectious-diseases animal contact diseases	-0.6001	-0.5683	-0.6493	-0.599	-0.6282	-0.5924	All coefficients uniformly indicate strong negative correlation. A logical assumption considering that such diseases are abundant in developing countries which do not have the necessary educational funding and capabilities.
major-infectious-diseases water contact diseases	-0.5705	-0.5555	-0.6028	-0.569	-0.5842	-0.5651	Same as above.
~~environment-international-agreements party to~~	-	-	-	-	~~0.5009~~	~~-0.4643~~	Considering that coefficients could not be calculated with confidence ($p<0.05$) for male and female it would be dangerous to

							generalize thus this observation is discarded.
dependency-ratios potential support ratio *"the number of working-age people (ages 15-64) per one elderly person (ages 65+)"*	-0.4345	-0.7436	-0.4427	-0.7688	-0.4491	-0.7678	Extremely interesting finding which could constitute the null hypothesis of a future paper. The difference between Pearson's and Spearman's coefficients is justified by the fact that the dependency ratios column is not normally distributed thus the Spearman's correlation coefficient is more accurate. Overall, it seems that as the elderly increase in a society (dependency support ratio decreases) the higher education quality gets. This may be justified by the fact that developed countries have a birth deficit relatively to developing ones.
urbanization rate of urbanization	-0.6478	-0.6414	-0.6886	-0.6747	-0.6712	-0.6708	Coefficients strongly suggest negative correlation. If the urbanization trend in a country increases, education quality decreases.
infant-mortality-rate total	-0.7367	-0.8102	-0.8179	-0.8386	-0.7855	-0.8464	Very strong negative correlation. The higher the infant mortality rate, the lower the quality of education offered. Once again, developing countries are targeted.
infant-mortality-rate male	-0.7425	-0.8106	-0.8218	-0.8384	-0.7906	-0.8452	
infant-mortality-rate female	-0.7266	-0.8082	-0.8099	-0.836	-0.7761	-0.8447	
drinking- unimproved-water-source urban	-0.3107	-0.6146	-0.3464	-0.6134	-0.3313	-0.6344	All three water-related entries refer to the access in unimproved (in terms of sanitation) water resources. That means that the lack of access to clean water resources means poorer education.
drinking- unimproved-water-source rural	-0.6507	-0.7122	-0.707	-0.7335	-0.6858	-0.7384	
drinking-unimproved-water-source total	-0.6294	-0.7222	-0.675	-0.739	-0.6573	-0.7476	
unimproved-sanitation-facility-access urban	-0.6795	-0.7301	-0.7103	-0.7454	-0.6953	-0.7498	As described above for water, the same applies for sanitation facilities. Lack of sanitation is highly correlated with poor education, probably due to developing countries.
unimproved-sanitation-facility-access rural	-0.7186	-0.7375	-0.767	-0.7718	-0.7478	-0.77	
unimproved-sanitation-facility-access total	-0.7211	-0.7493	-0.7669	-0.7788	-0.7457	-0.7775	
hiv-aids-deaths	-0.2558	-0.5291	-0.2634	-0.5504	-0.2468	-0.5562	The HIV-deaths column is not normally distributed thus Spearman's correlation

							coefficient is better in explaining the data. Overall, it seems that a smaller correlation exists compared to the sanitation and water access fields. The lower correlation between quality of education and HIV deaths versus water and sanitation access may be explained by the fact that developed countries are also facing HIV deaths while they enjoy full sanitation and zero unimproved water access.
gdp-per-capita-purchasing power parity *"GDP on a purchasing power parity basis divided by population"*. Basically GDP per capita considering the cost of living in the US.	0.5249	0.7081	0.511	0.7216	0.5311	0.7362	A worrying discrepancy between the coefficients is observed. Considering that the GDP data are normally distributed, Pearson's coefficient is more appropriate for the analysis of this bivariate correlation. A correlation does exist but I would refrain from characterizing it strong. High GDP per capita along with a strong currency seem to be attributes of countries with high quality education.
gdp-composition-by-sector-of-origin agriculture	-0.6086	-0.6069	-0.6458	-0.6486	-0.6317	-0.6385	Agreement between coefficients and values <-0.5 suggest strong negative correlation. Agricultural-based economies tend to have poorer education and vice-versa (determining which variable is independent requires further research which falls outside the scope of this paper).
~~gdp-composition-by-sector-of-origin services~~	-	-	~~0.4993~~	~~-0.5266~~	~~0.4821~~	~~-0.5113~~	This observation is discarded for three reasons: 1) Lack of statistical significance for "Male" school life expectancy correlation 2) Relatively moderate coefficients 3) Similarity with the "labor-force-by-occupation services" data which provide higher confidence coefficients.

labor-force-by-occupation agriculture	-0.7064	-0.6834	-0.7423	-0.699	-0.7366	-0.7072	Unarguably a strong negative correlation exists between the amount of people employed in agriculture and quality of education in a country. The opposite correlation (positive) but weaker is observed for employment in services.
labor-force-by-occupation services	0.5469	0.5185	0.5783	0.5484	0.5694	0.5443	
population-below-poverty-line	-0.5731	-0.5985	-0.6136	-0.6301	-0.5968	-0.6368	An expected correlation. The percentage of people below the poverty line in a society has an opposite (negative) relation with quality of education. As the people below the poverty line increase, education quality decreases.
~~central-bank-discount-rate~~	-	-	-	-	~~-0.3782~~	~~-0.5044~~	Lack of agreement and confidence of coefficients leads to the rejection of this observation.
commercial-bank-prime-lending-rate *"compares a simple average of annualized interest rates commercial banks charge on new loans, denominated in the national currency, to their most credit-worthy customers".* How safe is a country to lend money to. The smaller the better.	-0.403	-0.5488	-0.4071	-0.5508	-0.4133	-0.5687	This rate describes the stability and security of an economy. Consequentially, as the rate increases the lower the confidence placed on the economy is. Weak and unstable economies are correlated with lower quality education.
~~stock-of-domestic-credit~~	~~0.1751~~	~~-0.5077~~	-	-	~~0.1734~~	~~-0.5089~~	Considering that the stock of credit data are normally distributed data the Pearson's correlation is too small to draw safe conclusions hence this observation is discarded.
~~exchange-rates~~	-	-	~~-0.1325~~	~~-0.5481~~	-	-	Lack of high confidence coefficients for male and total school life expectancy along with discrepancies in available coefficients cannot guarantee safe assumptions.
electricity-access population without electricity	-0.3049	-0.709	-0.3261	-0.7091	-0.3159	-0.7164	All three electricity-related sets of data are not normally distributed thus more confidence is placed on the Spearman's correlation coefficients. Nonetheless, there is overall agreement that electrification is
electricity-access electrification - total population	0.6824	0.7117	0.7301	0.7102	0.7124	0.7237	
electricity-access electrification - rural areas	0.6428	0.6341	0.7197	0.7053	0.6906	0.6803	

							strongly and positively correlated to education.
~~natural-gas-imports~~	~~0.2908~~	~~0.5037~~	~~0.2444~~	~~0.5115~~	~~0.2737~~	~~0.5214~~	The natural gas imports data set is normally distributed but the Pearson's coefficient remains low. Moreover, Spearman's coefficient is strangely larger. After further visual inspection it was determined that correlation cannot be safely established.
telephones-fixed-lines subscriptions per 100 inhabitants	0.567	0.6867	0.5588	0.7019	0.5647	0.7099	The distribution of data is not normal thus Spearman's coefficient is more appropriate. High positive correlation, probably influenced by the fact that developed countries have way higher amounts of landlines as a percentage of population.
~~telephones-mobile-cellular subscriptions per 100 inhabitants~~	-	-	~~0.4447~~	~~0.5029~~	-	-	Lack of confidence in explaining this correlation.
internet-users percent of population	0.7111	0.7014	0.747	0.7316	0.7426	0.7385	Strong positive correlation. Contrary to common belief, internet usage remains low in certain developing countries. 101 out of 238 Factbook entries (countries, territories, disputed areas) have less than 50% of their population with access to the internet.
~~national-air-transport-system annual passenger traffic on registered air carriers~~	~~0.2136~~	~~0.5011~~	-	-	-	-	Assumptions cannot be made with confidence not only due to the lack of coefficients for female and total but also because of discrepancy in the existing values.
hospital-bed-density The amount of hospital beds per 1000 people.	0.4148	0.5667	0.4199	0.5844	0.4234	0.5887	The bed density data do not follow a normal distribution thus Spearman's correlations is more accurate. A moderate to strong correlation exists between quality of education and healthcare. The wider the health infrastructure, the stronger education tends to be. Interestingly, this

							correlation is not as strong as others, for example internet usage.
obesity-adult-prevalence-rate	0.4175	0.5042	0.4712	0.5472	0.4397	0.5269	The bivariate positive correlation between obesity rates and quality of education is attributed to the fact that obesity occurs in developed countries which also have comparatively stronger education than developing countries. What is interesting is the strength of the correlation when compared with other factors in this list like hospital beds per population with which they carry almost equal weight.
broadband-fixed-subscriptions per 100 inhabitants	0.7079	0.7679	0.689	0.7704	0.7073	0.7834	Strong positive correlation.
maternal-mortality-rate	-0.6673	-0.8252	-0.7221	-0.8541	-0.7002	-0.8532	Strong negative correlation. Maternal mortality rate is almost null in developed countries while high in developing.
physicians-density	0.6327	0.7037	0.6629	0.7427	0.6587	0.7381	Once again, indicative and related with developed versus developing countries.
stock-of-direct-foreign-investment-at-home	0.3572	0.5517	0.3041	0.5358	0.3439	0.5514	Generally, strong economies tend to have strong education. Discrepancies in data and moderate positive correlation observed.
stock-of-direct-foreign-investment-abroad	0.3354	0.5295	0.2807	0.5088	0.3288	0.5269	
contraceptive-prevalence-rate *"the percent of women of reproductive age (15-49) who are married or in union and are using, or whose sexual partner is using, a method of contraception according to the date of the most recent available data. The contraceptive prevalence rate is an indicator of health services, development, and women's empowerment."*	0.614	0.6145	0.6976	0.6865	0.6698	0.6649	As an indicator of "*health services, development, and women's empowerment*" the prevalence rate does an excellent job in describing how developed a country is in a condensed manner. All of the components of a society expressed in this rate are also directly related with education. This rate can be useful for future classification tasks.
children-under-the-age-of -5--years-underweight	-0.6976	-0.7273	-0.7411	-0.7403	-0.7274	-0.7355	Strong negative correlation. Underweight children, due to malnutrition is a characteristic of developing countries whose resources and living standards do not suffice for proper nutrition. Thus,

							education is neglected for the sake of survival.
major-infectious-diseases degree of risk_none	0.5599	0.5559	0.5948	0.5878	0.5879	0.588	The non-existence of major infectious diseases is positively correlated with quality of education. An equal but positive trends exists for the existence of such diseases.
major-infectious-diseases degree of risk_very high	-0.5516	-0.5381	-0.5986	-0.573	-0.5803	-0.5691	
Region_africa	-0.5885	-0.563	-0.6195	-0.5854	-0.6059	-0.5717	These correlations are the heart of this correlations table. The dichotomy between developed and developing countries has been repeated numerous times in the comments column of this table. This dichotomy is materialized here in a raw manner. European countries are positively correlated with quality education while the opposite happens for African ones. These correlations reveal something more important though. The absence of other continents from this table reveals something more important. It reveals the existence of homogeneity in the education of Europe and Africa which extends to homogeneity in development.
Region_europe	0.4883	0.5294	0.4743	0.5158	0.4897	0.5386	
age-structure - 25-54 years			0.5115	0.4844			Due to the lack of significant correlation for male and total caution is advised in the interpretation of the female one. Are aging societies mostly developed ones? Are women excluded from education in developing countries (which are not aging)? Even though these two hypotheses require proof in an academic setting, I would dare to accept them as universal realities. If both of these hypotheses are accepted then this correlation can be explained. The 25-54 age structure is higher in developed countries with aging populations. Such societies offer

							more opportunities for women than developing ones do. It should be noted that this assumption lacks sufficient proof in academic terms but could be accepted as a universal reality. Caution is advised, further analysis suggested.
sex-ratio at birth *"the number of males for each female"*	0.3866	0.5288			0.3691	0.5234	Sex ratio at birth is normally distributed thus Pearson's coefficient is more appropriate. Small correlation is suggested with low confidence considering that female coefficients were $p > 0.05$. Does gender composition correlate with quality of education? The lack of sex ratios for ages 15+ makes this assumption unsafe to make. Further analysis is suggested in future paper.
sex-ratio – 0-14 years	0.41	0.5366	0.3649	0.5015	0.3971	0.5351	
life-expectancy-at-birth total population	0.7831	0.7912	0.8261	0.8091	0.8087	0.8113	Life expectancy codifies a plethora of factors. Countries with high life expectancy are strongly correlated with strong educational systems.
life-expectancy-at-birth male	0.7722	0.7655	0.8121	0.7772	0.7964	0.7831	
life-expectancy-at-birth female	0.7852	0.8118	0.8302	0.8329	0.8119	0.8353	
literacy male	0.6899	0.7389	0.7486	0.7381	0.7313	0.7599	This is one of the few cases where causation and direction of correlation can also be established. Education leads to higher literacy rates.
literacy female	0.6882	0.7022	0.7607	0.7164	0.7357	0.7284	
mother-s-mean-age-at-first-birth	0.7774	0.7982	0.8	0.8154	0.8006	0.8254	Very strong positive correlation between women's age of first birth and quality of education. Societies in which women give birth at older ages have higher quality of education. Once again indicative of developed versus developing countries.
dependency-ratios total dependency ratio *"the ratio of combined youth population (ages 0-14) and elderly population (ages 65+) per 100 people of working age (ages 15-64). A high total dependency ratio indicates that the working-age population and*	-0.5929	-0.4851	-0.6573	-0.5172	-0.6366	-0.5317	The size of non-productive age groups in a country negatively affects the quality of education. Interestingly, considering that birth deficit and aging occurs in developed countries, the highest dependency ratios

the overall economy face a greater burden to support and provide social services for youth and elderly persons, who are often economically dependent."							are observed in developing countries probably due to the fact that the younger population is younger than 14 thus accounted as nonproductive.
median-age total	0.7653	0.7916	0.7852	0.8102	0.783	0.8203	The median age is directly linked with the life expectancy which is also part of the list of correlations. Countries with high living standards and strong healthcare also have strong educational systems (see the Scandinavian example).
median-age male	0.7686	0.7906	0.7876	0.8053	0.7855	0.8182	
median-age female	0.7581	0.7884	0.7785	0.8085	0.7765	0.8176	
literacy total population	0.6907	0.7153	0.7597	0.7228	0.7371	0.7398	Directly linked with quality of education. The effect of a strong educational system.
age-structure -55-64 years	0.7141	0.7385	0.7348	0.7623	0.735	0.7696	Directly linked with high median ages, life expectancy and birth deficit. Indicative of developed countries.
population-growth-rate	-0.5148	-0.5607	-0.5535	-0.5845	-0.5366	-0.5834	Population growth is related to weak educational systems.
age-structure – 0-14 years	-0.7445	-0.7453	-0.7825	-0.7673	-0.7713	-0.7755	Directly linked with low median ages, life expectancy and birth surplus. Indicative of developing countries. The stalemate occurring in developing countries is illustrated. The large amount of youth cannot receive adequate education leading to lack of development which means limited resources. This negative loophole is increasingly difficult to break.
age-structure -15-24 years	-0.718	-0.7633	-0.7205	-0.7683	-0.7244	-0.7836	
dependency-ratios youth dependency ratio *"youth dependency ratio - The youth dependency ratio is the ratio of the youth population (ages 0-14) per 100 people of working age (ages 15-64)."*	-0.7432	-0.7355	-0.7914	-0.7574	-0.7783	-0.7751	The Factbook itself provides with the justification: *"A high youth dependency ratio indicates that a greater investment needs to be made in schooling and other services for children."*. When combined with struggling developing countries, poor education occurs.
birth-rate	-0.7342	-0.7511	-0.7859	-0.7741	-0.7674	-0.781	These rates are higher in developing countries which have poorer educational
total-fertility-rate	-0.6652	-0.676	-0.7281	-0.7027	-0.7043	-0.7092	

							systems. Moreover, people work from young ages to preserve their families.
age-structure years 65 and over	0.7077	0.7742	0.7126	0.7937	0.7185	0.802	Directly linked with high median ages, life expectancy and birth deficit. Indicative of developed countries.
dependency-ratios elderly dependency ratio *"elderly dependency ratio - The elderly dependency ratio is the ratio of the elderly population (ages 65+) per 100 people of working age (ages 15-64)."*	0.6922	0.76	0.6893	0.7727	0.6961	0.7786	
electricity-access electrification - urban areas	0.5278	0.4304	0.5786	0.4635	0.5543	0.4453	The concept of electricity has been addressed above in the other related fields.
All correlations presented are $p < 0.05$. For $p > 0.05$ coefficients were rejected and cells were left empty.							

The Correlation analysis of Group 2 "education-expenditures" did not yield any strong correlation (coefficient > 0.5) and only few expected (related to GDP and taxation) ones.

Summarizing the findings of the correlation analysis, the following major conclusions can be reached:

1. Life expectancy (also expressed through median age, society age structure and dependency ratios) has the highest positive correlation. It is probably the most heavily correlated factor with quality of education. Countries with high life expectancy and old population tend to have higher quality education than those with more youth than elders.
2. Financial indicators are largely absent from the list of correlations and those present occupy intermediate positions in terms of correlation intensity. Social factors seem to be more important while one may argue that economic elements are embedded in other data like healthcare and life expectancy.
3. Infant mortality rate and sanitation are the most negatively correlated factors.
4. Women are present in various correlation factors. From birth rates to prevalence rates and maternal mortality rate.
5. Out of all continent labels, only Africa and Europe made it into the correlations list. This means that each one of these two continents are more homogenous in terms of education compared to every other continent. Specifically, Africa has a strong negative correlation with quality education while Europe a strong positive one.

Clustering

Considering that education is approached in a qualitative manner, the column "education-expenditures" was not used in the clustering process since it does not adequately describe the quality of education in a country, as illustrated in The Data section. Moreover, UN's Education Index (UNDP, 2018) was also appended to the CIA World Factbook dataset and used as feature along the school life expectancy data. Firstly, MeanShift clustering was conducted but no clusters were found. Moving on, the elbow and silhouette index methods were used to identify the optimal number of clusters.

Figure 6 - Elbow method for optimal number of clusters.

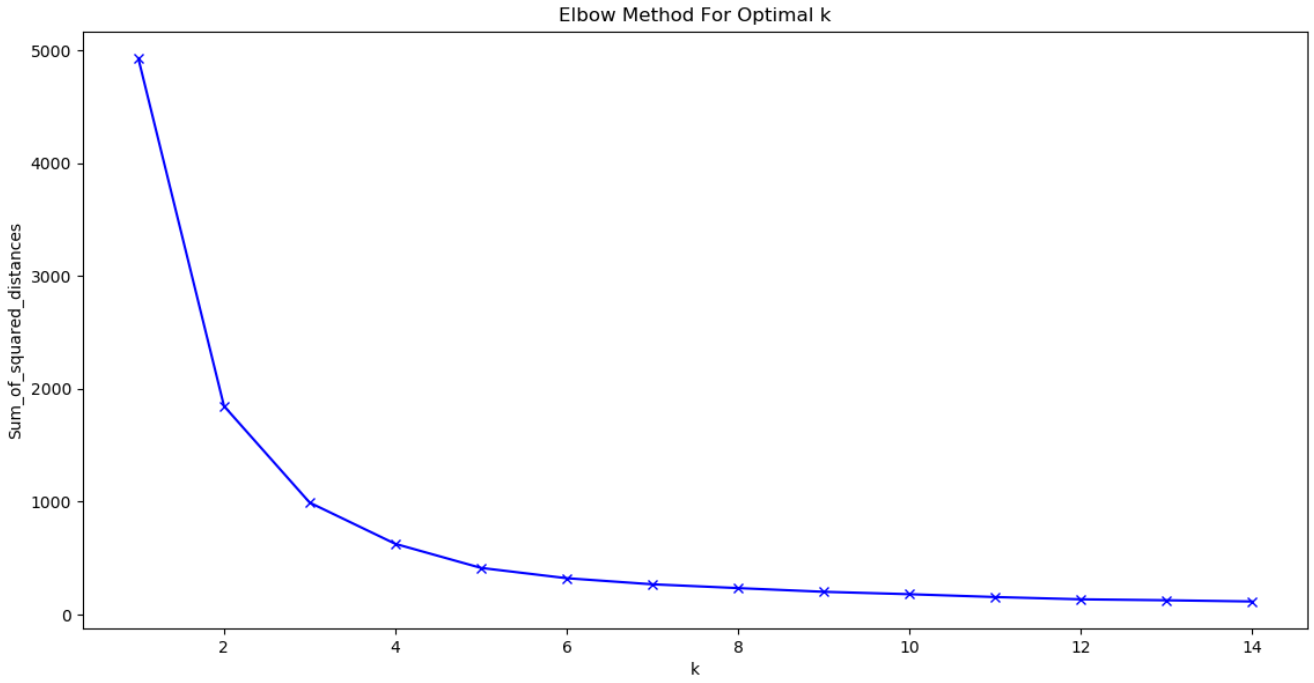
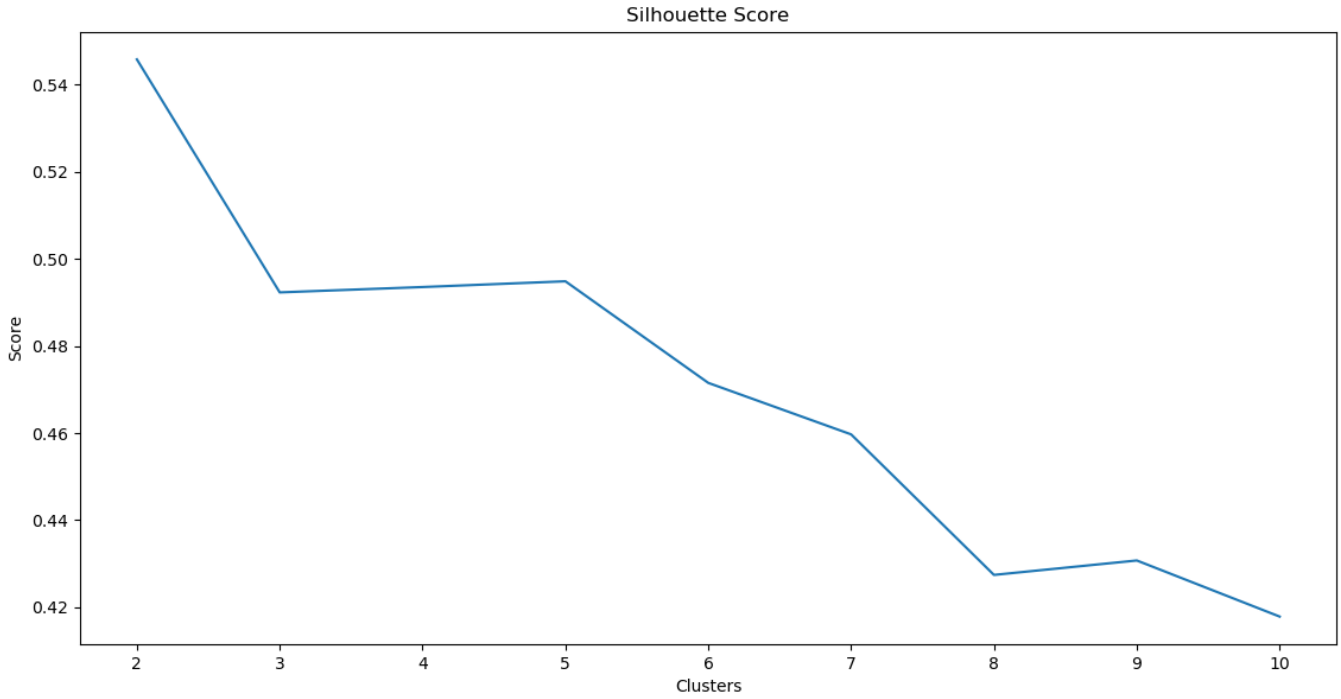


Figure 7 – Silhouette method (index) for optimal number of clusters.



The optimal number of clusters was determined to be 2. KMeans clustering was then performed on the data with: School life expectancy male, School life expectancy female, School life expectancy total, UN education Index Score being the features. 179 countries were classified (rest dropped because of missing data).

Table 4 - Clustering Metrics.

KMeans Clustering Model Metrics	
Inertia	1878
Davies Bouldin Score	0.61
Silhouette Score	0.54

The fact that MeanShift failed to effectively cluster the observations along with the moderate metrics raises concerns about the validity of clustering results.

Table 5 - Clustering Outcomes.

	Cluster (Group) 1	Cluster (Group) 2
Members	Antigua And Barbuda, United Arab Emirates, Algeria, Azerbaijan, Albania, Armenia, Andorra, American Samoa, Argentina, Australia, Austria, Bahrain, Barbados, Botswana, Belgium, Bahamas, Belize, Bosnia And Herzegovina, Bolivia, Belarus, Solomon Islands, Brazil, Bhutan, Bulgaria, Canada, Sri Lanka, China, Chile, Colombia, Costa Rica, Cuba, Cabo Verde, Cyprus, Denmark, Dominica, Dominican Republic, Ecuador, Egypt, Ireland, Estonia, El Salvador, Finland, Fiji, Micronesia, France, Georgia, Grenada, Germany, Greece, Hong Kong, Croatia, Hungary, Iceland, Indonesia, Iran, Israel, Italy, Iraq, Japan, Jamaica, Jordan, Kyrgyzstan, Kuwait, Kazakhstan, Latvia, Lithuania, Slovakia, Liechtenstein, Luxembourg, Libya, Mongolia, Montenegro, Mauritius, Malta, Oman, Maldives, Mexico, Malaysia, Vanuatu, Netherlands, Norway, Suriname, Nicaragua, New Zealand, Peru, Poland, Panama, Portugal, Palau, Qatar, Serbia, Romania, Philippines, Saudi Arabia, Saint Kitts And Nevis, Seychelles, South Africa, Slovenia, Singapore, Spain, Saint Lucia, Sweden, Switzerland, Trinidad And Tobago, Thailand, Tonga, Sao Tome And Principe, Tunisia, Timor-Leste, Turkey, United Kingdom, Ukraine, United States, Uruguay, Saint Vincent And The Grenadines, Venezuela, Samoa	Afghanistan, Angola, Bangladesh, Benin, Burundi, Cambodia, Chad, Congo, DR Congo, Cameroon, Comoros, Central African Republic, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gambia, Gabon, Ghana, Guatemala, Guinea, Guyana, Haiti, Honduras, India, Kenya, Kiribati, Laos, Lebanon, Liberia, Lesotho, Madagascar, Moldova, Malawi, Mali, Morocco, Mauritania, Mozambique, Niger, Nigeria, Nepal, South Sudan, Paraguay, Pakistan, Papua New Guinea, Guinea-Bissau, Rwanda, Senegal, Sierra Leone, Somalia, Sudan, Tajikistan, Togo, Turkmenistan, Tanzania, Uganda, Burkina Faso, Uzbekistan, Namibia, Yemen, Zambia, Zimbabwe
Number of Countries	117	62
Avg. UN Education Score (2015)	0.73 (+/- 0.11)	0.45 (+/- 0.11)
Avg. Total SLE	15 (+/- 1.8)	10 (+/- 1.8)
Avg. Female SLE	15 (+/- 2)	9.7 (+/- 2)
Avg. Male SLE	14 (+/- 1.7)	10.3 (+/- 1.7)
Avg. Education Expenditures	5 (+/- 1.9) % of GDP	4 (+/- 1.9) % of GDP

Group 1 represents countries with strong educational systems while Group 2 the opposite. The clustering, at first, did not reveal any important information. Although, when the average education expenditures for each cluster were calculated, it became evident that differences are minor. This can be attributed to the fact that education expenditures are expressed as percentages of GDP. In that regard, countries with small per capita GDPs are unable to provide quality of education regardless of the percentage of GDP invested in education due to the fact that the GDP itself is small.

It should be further noted that Group 1 has a higher average for female compared to male. The opposite trend is observed in Group 2. Countries with quality education provide opportunities for women while countries with poorer education do not.

Various developing countries made it into Group 1. This means that quality education is not a luxury of developed nations only.

Conclusion & Ideas for further research.

The correlations analysis and clustering led to various important conclusions all listed below:

1. While commenting on the bivariate correlations, the developed versus developing country dichotomy kept repeating. Further research on whether education can be a key feature for the classification of countries into developed and developing ones is suggested. While the correlations revolved around the developed – developing dichotomy, clustering did not seem to prove it.
2. There is a general lack of education-related qualitative metrics hence the failure of clustering algorithms. Can the quality of education be calculated indirectly through strongly correlated data?
3. European countries offer the highest quality education and African states the lowest, homogeneity considered.
4. Education is a complex multi-correlated phenomenon. Closely related with life expectancy and sanitation. Lower and fewer correlations with financial factors.
5. Participation of women in education is a strong measure of educational system quality.
6. The clustering algorithm showed inability to adequately group countries. This leads to the assumption that countries either have diverse and unique educational characteristics which could be better studied on a “per case” basis or that existing measures (like School Life Expectancy and UN’s, PISA scores) are unable to provide with an accurate depiction of educational realities.
7. More attention should be given to the correlations section of this paper and not on clustering.
8. Healthcare, sanitation and social factors are more correlated with education than economic ones.

Future work on the field may focus on causation between various factors and education or even a study between economic indexes and literacy.

References

CIA, 2019. *Notes and Definitions.* [Online]
Available at: https://www.cia.gov/library/publications/the-world-factbook/docs/notesanddefs.html#371
[Accessed 01 May 2020].

CIA, 2019. *Publications: Download.* [Online]
Available at: https://www.cia.gov/library/publications/resources/the-world-factbook/index.html
[Accessed 25 August 2019].

Eric, J., Travis, O., Pearu, P. & others, a., 2001. *SciPy: Open Source Scientific Tools for Python.* s.l.:s.n.

JetBrains, s., 2000. *JetBrains: Pycharm.* [Online]
Available at: https://www.jetbrains.com/pycharm/
[Accessed 29 August 2019].

Kouskouvelis, I., 2007. *Εισαγωγή στις Διεθνείς Σχέσεις.* Athens: POIOTHTA.

OECD, 2018. *PISA 2018 Results (Volume I): What Students Know and Can Do.* [Online]
Available at: https://www.oecd.org/pisa/publications/pisa-2018-results.htm
[Accessed 3 June 2020].

Podiotis, P., 2020. *Towards International Relations Data Science: Mining the CIA World Factbook.* Thessaloniki: s.n., https://arxiv.org/abs/2010.05640

Project, NLTK, n.d. *NLTK.* [Online]
Available at: https://www.nltk.org/index.html
[Accessed 18 May 2020].

Python Foundation, S., 2001. *Python.* [Online]
Available at: https://www.python.org/
[Accessed 20 August 2019].

Tennage, P., 2018. *stack overflow.* [Online]
Available at: https://stackoverflow.com/questions/37487830/how-to-find-probability-distribution-and-parameters-for-real-data-python-3
[Accessed 30 May 2020].

UNDP, 2018. *UNITED NATIONS DEVELOPMENT PROGRAMME.* [Online]
Available at: http://hdr.undp.org/en/data#
[Accessed 3 June 2020].

Unver, H. A., 2018. *Computational International Relations. What Can Programming, Coding and Internet Research Do for the Discipline?,* s.l.: s.n.