Data Analysis of Human Development Index (HDI)

Shehab Shahin

University of Sydney | September 2018

Table of Contents

# Executive Summary

This report aims to study and understand the differences in social and economic development and their impacts on countries. This report uses the Human Development Index (HDI) data and its components, provided by the United Nations Development Programme (UNDP), to evaluate the findings. The HDI data takes life expectancy at birth, expected and mean years of schooling, and the Gross National Income (GNI) for each country into account. When calculated, it provides a numeric value ranging between 0 (lowest) and 1 (highest) and also a ranking (ascending order), giving us 1 dependent variables (HDI), and 4 independent variables. In the report, 3 questions were proposed and answered using Rstudio to further clarify the data and assist in understating it.

Section one of the report focuses on the initial data analysis, in which the data is explained. By doing so, we give a summary of what the data is showing, classify variables, provide the source of the data, identify the stakeholders and domain knowledge.

Section two of the report focuses on answering 3 questions that provide more detailed conclusions about the data. The first question attempts to study the relationship between the mean years of schooling and the gross national income per capita. A scatter plot was used to study the correlation between the variables, in which it provided a positive correlation showing a close proportional relationship between the variables. The second question attempts to understand the difference in education expectations in order to assess the difference in education standards between countries as the HDI varies. By using a box and whisker plot and a bar plot, we were able to come up with the conclusion that there was little difference in the expected years of schooling. The third question examines the difference in life expectancy at birth between countries of low and high HDI. To provide the desired conclusion, a scatter plot and a boxplot were used in providing the answer. A clear difference was seen as the life expectancy tended to decrease towards the lower ranked countries.

# Full Report

## Initial Data Analysis (IDA):

setwd("E:/Work/GitHub Portfolio/Projects/Data Analysis of HDI Project")  
HDI = read.csv("2016\_Statistical\_Annex\_Table\_1.csv", header = T)

message("Structure of the data which consists of both quantitative and qualitative variables.")

## Structure of the data which consists of both quantitative and qualitative variables.

str(HDI)

## 'data.frame': 188 obs. of 9 variables:  
## $ HDI.rank : int 1 2 2 4 5 5 7 8 9 10 ...  
## $ Country : chr "Norway" "Australia" "Switzerland" "Germany" ...  
## $ Human.Development.Index..HDI. : num 0.949 0.939 0.939 0.926 0.925 0.925 0.924 0.923 0.921 0.92 ...  
## $ Life.expectancy.at.birth : num 81.7 82.5 83.1 81.1 80.4 83.2 81.7 81.1 82.7 82.2 ...  
## $ Expected.years.of.schooling : num 17.7 20.4 16 17.1 19.2 15.4 18.1 18.6 19 16.3 ...  
## $ Mean.years.of.schooling : num 12.7 13.2 13.4 13.2 12.7 11.6 11.9 12.3 12.2 13.1 ...  
## $ Gross.national.income..GNI..per.capita: num 67614 42822 56364 45000 44519 ...  
## $ GNI.per.capita.rank.minus.HDI.rank : int 5 19 7 13 13 -3 8 11 20 12 ...  
## $ HDI.rank.1 : int 1 3 2 4 6 4 6 8 9 9 ...

D = dim(HDI)  
message("Dimensions of the data are: ", toString(D))

## Dimensions of the data are: 188, 9

names(HDI)

## [1] "HDI.rank"   
## [2] "Country"   
## [3] "Human.Development.Index..HDI."   
## [4] "Life.expectancy.at.birth"   
## [5] "Expected.years.of.schooling"   
## [6] "Mean.years.of.schooling"   
## [7] "Gross.national.income..GNI..per.capita"  
## [8] "GNI.per.capita.rank.minus.HDI.rank"   
## [9] "HDI.rank.1"

head(HDI)

## HDI.rank Country Human.Development.Index..HDI. Life.expectancy.at.birth  
## 1 1 Norway 0.949 81.7  
## 2 2 Australia 0.939 82.5  
## 3 2 Switzerland 0.939 83.1  
## 4 4 Germany 0.926 81.1  
## 5 5 Denmark 0.925 80.4  
## 6 5 Singapore 0.925 83.2  
## Expected.years.of.schooling Mean.years.of.schooling  
## 1 17.7 12.7  
## 2 20.4 13.2  
## 3 16.0 13.4  
## 4 17.1 13.2  
## 5 19.2 12.7  
## 6 15.4 11.6  
## Gross.national.income..GNI..per.capita GNI.per.capita.rank.minus.HDI.rank  
## 1 67614 5  
## 2 42822 19  
## 3 56364 7  
## 4 45000 13  
## 5 44519 13  
## 6 78162 -3  
## HDI.rank.1  
## 1 1  
## 2 3  
## 3 2  
## 4 4  
## 5 6  
## 6 4

tail(HDI)

## HDI.rank Country Human.Development.Index..HDI.  
## 183 183 Guinea 0.414  
## 184 184 Burundi 0.404  
## 185 185 Burkina Faso 0.402  
## 186 186 Chad 0.396  
## 187 187 Niger 0.353  
## 188 188 Central African Republic 0.352  
## Life.expectancy.at.birth Expected.years.of.schooling  
## 183 59.2 8.8  
## 184 57.1 10.6  
## 185 59.0 7.7  
## 186 51.9 7.3  
## 187 61.9 5.4  
## 188 51.5 7.1  
## Mean.years.of.schooling Gross.national.income..GNI..per.capita  
## 183 2.6 1058  
## 184 3.0 691  
## 185 1.4 1537  
## 186 2.3 1991  
## 187 1.7 889  
## 188 4.2 587  
## GNI.per.capita.rank.minus.HDI.rank HDI.rank.1  
## 183 4 182  
## 184 5 184  
## 185 -8 185  
## 186 -19 186  
## 187 1 187  
## 188 4 188

## Classification of Variables:

There are 5 variables present, Human Development Index (HDI), Life expectancy at birth, Expected years of schooling, Mean years of schooling, Gross National Income (GNI) per capita. HDI is the independent variable and provides a ranking for the countries. Life expectancy at birth, expected years of schooling, mean years of schooling and GNI are all continuous, numerical data. Other data present that are not variables are GNI per capita rank minus HDI rank and HDI rank (2014). Both of these are discrete, numerical data.

## Source of Data:

The three data sources were retrieved from UNESCO, World bank and the United Nation. These are international reputable organizations, assessed and financed by a large number of governments and agencies. Therefore, the data that was provided by those organization is reliable as they are a result of primary investigations and experiments conducted by reputable individuals from different background. Also, the data is constantly updated to keep up with the new changes that are occurring regularly in the world. The data is also valid because it represents the multiple variables that contribute to the indexing of human development. It is also accurate as values were calculated to multiple decimal points.

## Stakeholders:

A large sector of the society can benefit from the HDI (Human Development Index), for example, charity aid workers, hospitals, governments, insurance companies etc… The trends presented by the data are of a significant importance to the stakeholders as they provide them with valuable information that can be utilized as a means of improvement in different areas depending on their level of involvement. For example, governments can utilise the HDI rankings to understand which sector needs to be improved upon within their country, whereas hospitals and primary health sectors can try to improve their treatment methods and procedure if the “Life expectancy at birth” is low.

## Domain Knowledge:

The domain knowledge that is required to interpret the data is very basic and includes understanding HDI, the factors which contribute to HDI and the basic definition of each factor. The factors that contribute to HDI are Life expectancy at birth, expected and mean years of schooling and Gross National Income (GNI) per capita. HDI is an index that provides information about the development of the population in a specific country using the factors mentioned above. The first factor “life expectancy at birth” is an estimation of how long an individual will live based on the mortality rate on the year the individual was born at. “Expected years of Schooling” is another factor that contributes to HDI and it’s also an estimation whereby the enrolment-rate is to be fixed throughout an individual’s life for it to be valid. “Gross national income (GNI) per capita” is the overall income of a nation measured in international dollars using PPP rates and divided by the midyear population. Finally, “mean years of schooling” is the average number of schooling years undertaken by individuals under 25 years old.

### Summary:

*The data came from The United Nations during their research into the HDI (Human Development Index) of all the countries/territories back in 2016.* Possible issues include the lack of data for some countries due their governments (i.e. North Korea) or countries/territories that are too small and the variables are unable to categorise them and thus do not have a HDI. Various columns also did not contain any data and thus were omitted as they skewed the data and changed the way R processed it (e.g. as a factor rather than num). *Each row represents the different countries/territories of the world, sorted in the order of highest HDI to lowest HDI in 2016* Each column represents the different variables used to rank each country’s HDI, those being “Human Development Index (HDI),” “Life Expectancy,” “Expected years of schooling,” “Mean years of schooling,” “Gross National Income (GNI) per capita,” “GNI per capita minus HDI rank,” and “HDI rank (2014).”

## Is there a relationship between the Mean Years of Schooling and the Gross National Income per capita?

The majority of a country/territory’s Gross National Income (GNI) comes from their economy and workforce. The gradient of the line is 0.59, making it positive. This means that as the mean years of schooling increases, the GNI of that nation also increases. The opposite can also be said, if the people of that country/territory do not have a high mean years of schooling, their country’s GNI is low. By comparing the mean years of schooling to the GNI, it can be established that a higher mean of years in school equates to a higher GNI of that respective country. Most countries are bound closely to the line of best fit however there were a few outliers. This also directly correlates to that country’s HDI. By also comparing the HDI to the GNI, another relationship is evident between the mean years of schooling and that country’s HDI. The gradient of the line for GNI vs HDIR is -0.78, meaning that as GNI decreases, there is a linear relationship to that of HDIR which will also decrease. Countries who have a higher GNI also have a higher rank. Most of the countries are close to the line of best fit, with fewer outliers in this comparison compared to the previous.

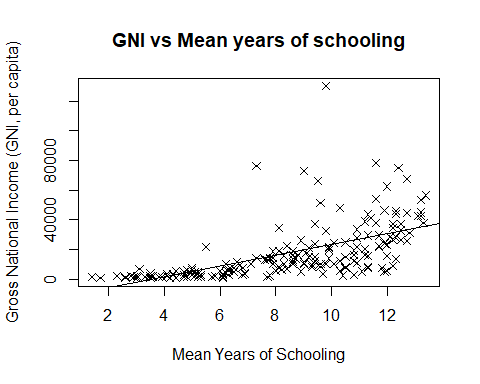
mschool = HDI$Mean.years.of.schooling  
GNI = HDI$Gross.national.income..GNI..per.capita  
C = cor(mschool, GNI, use = 'pairwise')  
message("Gradient of the line is: ", toString(C))

## Gradient of the line is: 0.592084326232843

plot(mschool, GNI, xlab = "Mean Years of Schooling", ylab = "Gross National Income (GNI, per capita)", main = "GNI vs Mean years of schooling", pch = 4)  
L = lm(GNI ~ mschool)  
L$coeff

## (Intercept) mschool   
## -13194.963 3644.462

abline(L)



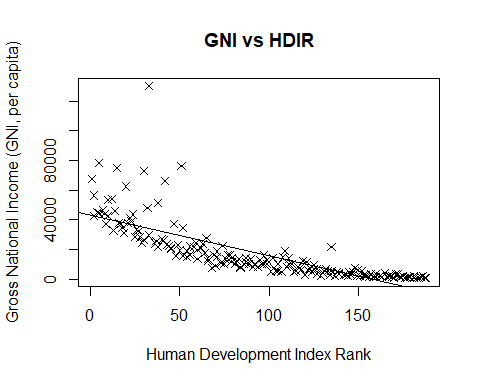
HDIR = HDI$HDI.rank  
GNI = HDI$Gross.national.income..GNI..per.capita  
CO = cor(HDIR, GNI, use = 'pairwise')  
message("Gradient of the line is: ", toString(CO))

## Gradient of the line is: -0.777828811030041

plot(HDIR, GNI, xlab = "Human Development Index Rank", ylab = "Gross National Income (GNI, per capita)", main = "GNI vs HDIR", pch = 4)  
L = lm(GNI ~ HDIR)  
L$coeff

## (Intercept) HDIR   
## 43014.0690 -272.4816

abline(L)



### Summary:

Summary: Overall there is a clear relationship between the 3 variables; mean years of schooling, gross national income (GNI) and human development index (HDI) rank. Mean years of schooling and GNI has a linear relationship that is proportional, whereby countries with a higher mean years of schooling tend to have a higher GNI than countries who have a lower mean years of schooling. Furthermore, the relationship between GNI and HDI rank proves to be linearly proportional as well: an increase of GNI means a higher ranking. Using these observations, it can be determined that a higher mean years of schooling and higher GNI will result in a high HDI rank.

## How much difference is there in the expected years of schooling?

The expected years of schooling is the number of years of schooling that a child of school entrance age can expect to receive if prevailing patterns of age-specific enrolment rates persist throughout the child’s life. From the numerical data, the average years that children around the world can expect to attend school is 12.95 years and the median is quite close at 13.10 years. This close relationship means that most children are expected to attend the same number of years of school. However, there is a large range of 15.50 years. The 1st quartile is at 10.9 years, meaning that bottom half of the population receive an average of 10.90 years of expected years of schooling. The 3rd quartile value is the average of the upper half of the population, situated at 15.00 years. Furthermore, the standard deviation of every country is a mere 2.86 years, meaning that most countries have a similar expected years of schooling. The bar plot shows a net decrease as the HDIR decreases, illustrating that a lower expected years of schooling results in a lower HDIR.

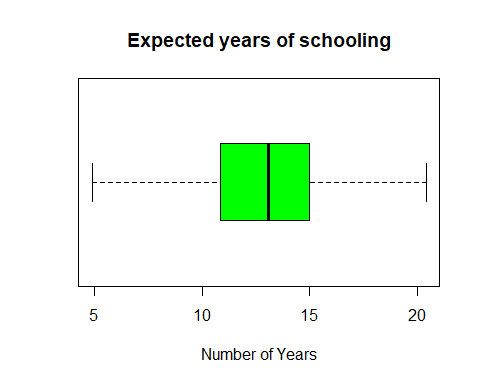
eschool = na.omit(HDI$Expected.years.of.schooling)  
s = sd(eschool)  
message("The standard deviation is: ", toString(s))

## The standard deviation is: 2.89526525063663

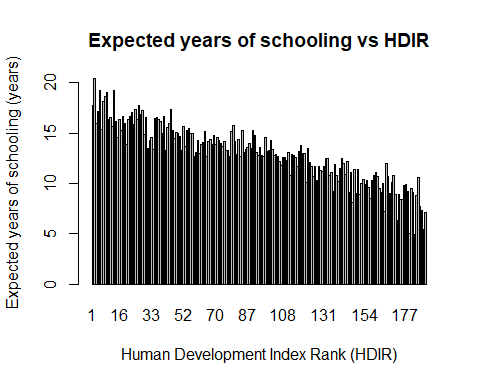
summary(eschool)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.90 10.88 13.10 12.98 15.00 20.40

boxplot(eschool, col = c("green"), main = "Expected years of schooling", xlab = "Number of Years", horizontal = T)



eschool <- na.omit(HDI$Expected.years.of.schooling)  
country <- na.omit(HDI$Country)  
barplot(HDI$Expected.years.of.schooling, names.arg = HDIR, xlab = "Human Development Index Rank (HDIR)", ylab = "Expected years of schooling (years)", main = "Expected years of schooling vs HDIR")

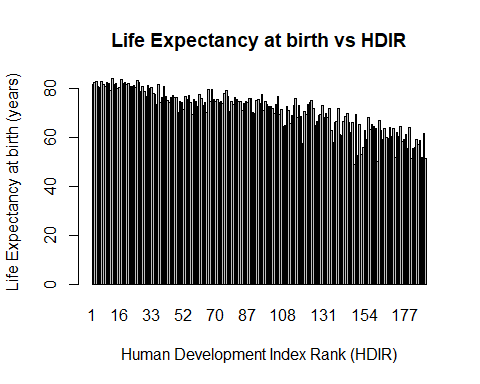


### Summary:

Summary: There is not much difference for the expected years of schooling across the world. The expected number years of schooling is quite similar throughout the world, with a median of 13.10 years of schooling whilst the mean is only just below at 12.98 years. Furthermore, the standard deviation is only 2.86, meaning that most countries are in close proximity of each other. Although the range is large, 15.60 years, there are no outliers because Q1 is 10.88 years and Q3 is 15.00 years. ## Does the life expectancy at birth differ a lot between high and low HDI countries?

Life expectancy is defined by the number of years a newborn is expected to for if the existing patterns of age-specific mortality rates stay constant throughout the infant’s life. Due to the large variation of countries and their status (i.e. at peace, at war, in poverty, etc), there is expected to be a large difference in the life expectancy. Looking at the bar plot, there is an obvious net decrease in the life expectancy as it goes towards the countries ranked with a low HDI. The median life expectancy is 73.40 years whilst the mean is 71.35 years. The range of life expectancy is 35.30 years with the highest being 84.20 and the lowest being 48.90 years. Furthermore, both graphs show a linear trend and the correlation for the top 20 and bottom 20 were quite high, thus proving there is a definite relationship between life expectancy and that country’s HDI ranking. The gradient for top 20 was -0.96 and the bottom 20 was -0.98. This correlates to the data correctly as countries with a lower life expectancy tended to also rank lower in terms of HDI than countries with a high life expectancy.

lifeexp = sort(na.omit(HDI$Life.expectancy.at.birth))  
barplot(HDI$Life.expectancy.at.birth, names.arg = HDIR, xlab = "Human Development Index Rank (HDIR)", ylab = "Life Expectancy at birth (years)", main = "Life Expectancy at birth vs HDIR")



summary(lifeexp)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 48.90 65.65 73.40 71.35 77.00 84.20

H = head(lifeexp, n=1)  
message("The shortest life expectancy is: ", toString(H))

## The shortest life expectancy is: 48.9

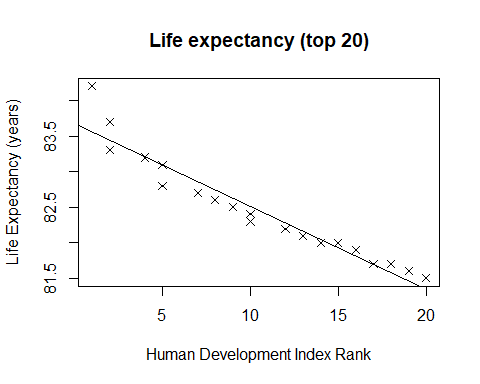
T = tail(lifeexp, n=1)  
message("The longest life expectancy is: ", toString(T))

## The longest life expectancy is: 84.2

par(mfrow = c(1,2))  
Co = cor(lifeexp[c(188:169)], HDIR[c(1:20)])  
message("Correlation between life expectancy and HDIR for top 20 countries is: ", toString(Co))

## Correlation between life expectancy and HDIR for top 20 countries is: -0.960782990594657

plot(HDIR[c(1:20)], lifeexp[c(188:169)], main = "Life expectancy (top 20)", xlab = "Human Development Index Rank", ylab = "Life Expectancy (years)", pch = 4)  
L = lm(lifeexp[c(188:169)] ~ HDIR[c(1:20)])  
abline(L)



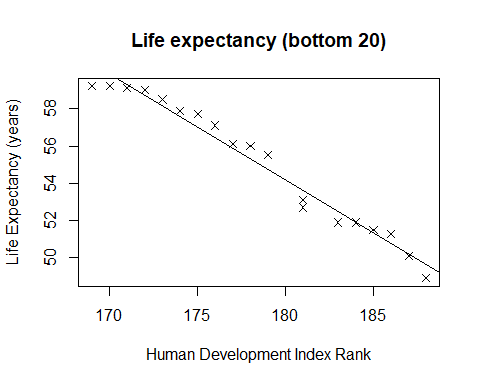
L$coeff

## (Intercept) HDIR[c(1:20)]   
## 83.6834118 -0.1167548

co = cor(lifeexp[c(1:20)], HDIR[c(188:169)])  
message("Correlation between life expectancy and HDIR for bottom 20 countries is: ", toString(co))

## Correlation between life expectancy and HDIR for bottom 20 countries is: -0.983658598841871

plot(HDIR[c(188:169)], lifeexp[c(1:20)], main = "Life expectancy (bottom 20)", xlab = "Human Development Index Rank", ylab = "Life Expectancy (years)", pch = 4)  
L = lm(lifeexp[c(1:20)] ~ HDIR[c(188:169)])  
abline(L)



L$coeff

## (Intercept) HDIR[c(188:169)]   
## 156.255847 -0.566961

### Summary:

Summary: The life expectancy at birth between the top and low HDI countries differ by a large margin. In both graphs we see a negative linear regression indicating a clear correlation between the variables whereby the decrease in HDI rank results in a decrease in life expectancy at birth. That means that people that live in countries with higher HDI are expected to live longer than ones with lower HDI. Furthermore, these observations allow us to conclude that the quality of the social and economic development has an influence on the life span of the people: The results allow us to assume that HDI is a forever changing statistic and external factors such as poverty, war, excessive diseases and others that are found in countries with low HDI shorten the life span of individuals.

# Conclusions

In conclusion, the statistical data was used as evidence that supported our answers to the research questions. Multiple variables are used in determining a country’s HDI and these variables are gathered by reputable sources such as UNESCO, World Bank and The United Nations. Research question number one showed that there is a trend between the three independent variables; mean years of schooling, gross national income (GNI) and human development index (HDI) rank and the trend was that if GNI or mean years of schooling increases then the HDI rank of the country will increase proportionally. The data was also used to answer the second research question and it was found that there is no significant difference in the expected years of schooling between the various countries in the world since the average was 13.10 with a standard deviation of only 2.90. Finally, the last research question showed that HDI has a high correlation with the life expectancy of different countries around the world. That is, individuals in developing countries and lower HDI can expect to have a shorter life expectancy than those in developed countries and high HDI countries.

# References

United Nations Development Programme. (2016). [online] Available at: <http://hdr.undp.org/en/composite/HDI> [Accessed 12 Sep. 2018].

Nationsonline.org. (2017). Human Development Index 2016 - HDI - Nations Online Project. [online] Available at: <https://www.nationsonline.org/oneworld/human_development.htm> [Accessed 13 Sep. 2018].