Analysis of Correlations between Human Development Index and IQ scores

Olle Rehnfeldt, Bence Pal, Christoph Kurzbauer, Jakob Hutter

rehnfeldt_olle@student.ceu.edu, pal_bence@student.ceu.edu, kurzbauer_christoph@student.ceu.edu, hutter_jakob@student.ceu.edu

Central European University

25th of February 2024

1 INTRODUCTION

This Anylsis investigates the correnation between the Human Development Index (HDI), a measure of a countries achievements in the field of human development, and countries average IQ score, an meaure of indivduals analytic intelligence. The Project in detail examines the variables interplay on a country level as well as on regional levels utiliting univariate and bivariate techniques, with focus on exploring the following Hypotheses:

1. The average IQ and Human Development Index are positively correlated on a Country Level.
2. Less developed regions are more prone to outliers.
3. The IQ is biased towards western countries.

1.1 Analysis Design

The project aims to answer the Hypotheses, and is structured as follows:

1. Cleaning, Preparation and Merging
2. Univariate for HDI and IQ Data separately
   On country and regional level
3. Bivariate Analysis for HDI and IQ

1.2 What the results can be used for

Firstly, the findings can be used to show that IQ is biased towards western country, which could be used to readjust the measurement norm in a way to reduce the bias,

Secondly, legislators can use the result to evaluate their country's IQ performance relative to others with similar HDI-Scores or within their respective region. The can further use these findings to readjust their strategies in terms of edcucation and other human development indicators that could have an infulence in IQ in the long-term.

1.3 Note on Causality

This project is an analysis of correlation not causality. We believe that the causality is that some components of the HDI are having a causal mechanism toward countries average IQ, and are keen to highlight that we don't believe there is causality in the reversed direction.

2 DATA

The Project combines two Databases, both data being from 2021. First, the Average IQ Score per country, the IQ Score is a normed measure of analytical intellgence of an individual. Second, the Human Development Index (HDI) per country, which is measure of countries achievements and standards in terms of health, education, and standard of living, the score ranges from one being the highest avhievable to zero being the lowest.

2.1 Variables

The two Datasets where merged into one Dataframe with the following columns (variables):

1. Country (Index)
2. Region
   from the HDI Datasource, relevant to investigate Hypothesis concerning local disparities.
   For simplification we will refer to them with their following abbreviations:
   - SA - South Asia
   - SSA - Sub Saharan Africa
   - ECA - Europe and Central Asia

    - AS - Rest of Asia
    - AME - Americas
    - EAP - East Asia and Pacific
3. HDI (Human Development Index 2021)
4. Average IQ (per country 2021)

Data Sources and Acquisition

The HDI Dataset is sourced from the United Nations Development Programme (UNDP), which is an UN development aid agency. The Data gets collected and published yearly. A possible bias for this dataset could be towards countries who are able to collect more data easier, this could be to financial advatages (western countries) or large population size, and further the dataset does not consider data transparecny, some countries could ommit or maniuplate data willingly.

The IQ Dataset is sourced from Kaggle and was published by Richard Lynn & Tatu Vanhanen, who since 2002 yearly publish IQ data as part of their book series “IQ and the wealth of Nations”. As mentioned in the Introduction Hypotheses section, we believe a bias for the IQ towards western countries exist, as analytical thinking is deeply embedded in the educational culutre, and the IQ is directed towards measuring exactly that.

For both IQ and HDI the available Data for the year 2021 was taken for this Project.

3 DATA CLEANING AND PREPARATION

First, we started by exploring the HDI dataset. It has 206 and 1008 columns and it contains the human development index of 194 countries from 1990 to 2021, the HDI rank of each country, the region the countries are located in, and the HDI code which shows whether the country has a low, medium, high or very high HDI. Other than that, more information is provided for each country which is irrelevant to our research (maybe mention the other data). After the exploration, we created a new HDI dataframe with the columns 'country', 'hdicode', 'hdi_2021', 'region' and 'hdi_rank_2021'. After renaming the columns, we checked the object type of each column and the countries that have no HDI assigned to them. These will then be removed because there is no way to fill in the HDI value for these. The original dataframe also contains the summary of HDI values for each region, which we save in a separate dataframe and then we drop these from the one we will work with so that we have only the countries. Then we remove all countries that have no HDI values assigned to them, but we keep those that have no region assigned to them. After that, we check for duplicated rows and we find none in the dataframe so we move on to check which columns contain nan values. We can see that 43 countries do not have a region assigned to them, so we check which countries these are exactly and we see that these are European, Asian, and North American countries. A unique region will then be assigned to each of these countries and a new one created for all countries from the American continent so that there is no separate North American region that only contains the USA and Canada. Then we sort the countries by the HDI in descending order and set the 'country' as the index which is important for the merging.

Then we explored and cleaned the IQ dataframe which was a lot cleaner than the HDI dataset. Again, we checked the columns, chose the columns of interest which were the countries and their respective average IQs, checked for duplicated rows and found none, and sorted the countries by their IQ values. We noticed that 41 countries had their IQ estimated from their neighboring countries. However, dropping these would have left us with an insufficient amount of countries for our analysis so we decided to keep them. At the end, we did the same as for the HDI dataframe and set the 'country' column as the index.

After that, we merged the two dataframes on the indexes. The method we used was the outer merge because we wanted to keep the countries from both dataframes so we could see which countries were missing from each. This helped us to investigate which countries were named differently in both dataframes which we then renamed in the original HDI and IQ dataframes. Finally, we merges the two dataframes with the renamed countries with an inner merge, which only kept the countries that were in both dataframes.

4 DESCRIPTIVE STATISTICS

4.1 Univariate analysis

To get an intial understanding of the distribution by, the data distirbution was plotted in a histogram as can be seen in Figure 1. Here, one can observe that the distribution of HDI has a a slight skew to the left, which is further strengthend by the skew statistic of -0.28. IQ is lessed skewed to the left with a score of -0.06, and ressembels more closely a normal distribution. However, IQ and HDIs distribution have a kurtosis socre of -0.83 and -0.91 respectively, which means that more values can be foun closer to the mean that would be expected in a normal distribution. It is also important to note that for population, IQ is centered around a 100 worldwide and us adjusted to keep it so, but since here each country is a data-entry and the popoulation of these countries can greatly differs, it is not centered at 100 but instead has a mean of 84.05. Looking at the

cumulative distributions, it can been seen that for HDI the slope of the function is pretty even, meaning that there is a similar amount of values in every score of the HDI. IQ, with its ressemblance to a normal distiribution is, more centered around the mean value which is seen by the steeper slope around the mean value in the function.

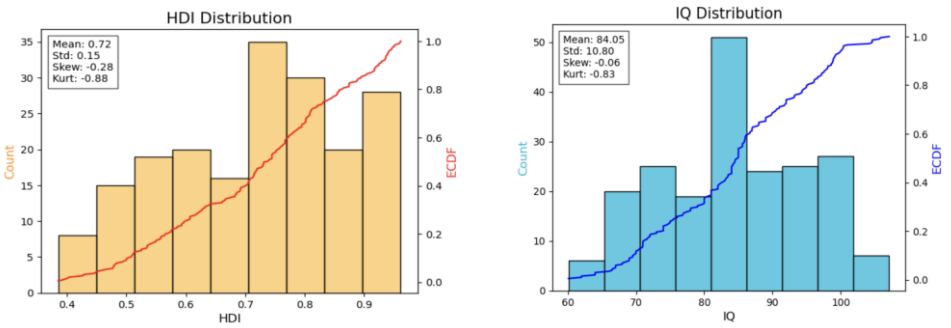

Figure 1

Determining if there were any outliers in the data was done with boxplots where data which was within the interquartile range was plotted inside the box and values which fell outisde the interquatile range multiplied by 1.5 was plotted outside. See Figure 2. This analysis was preformed regionally such that regional trends could be evaluated.

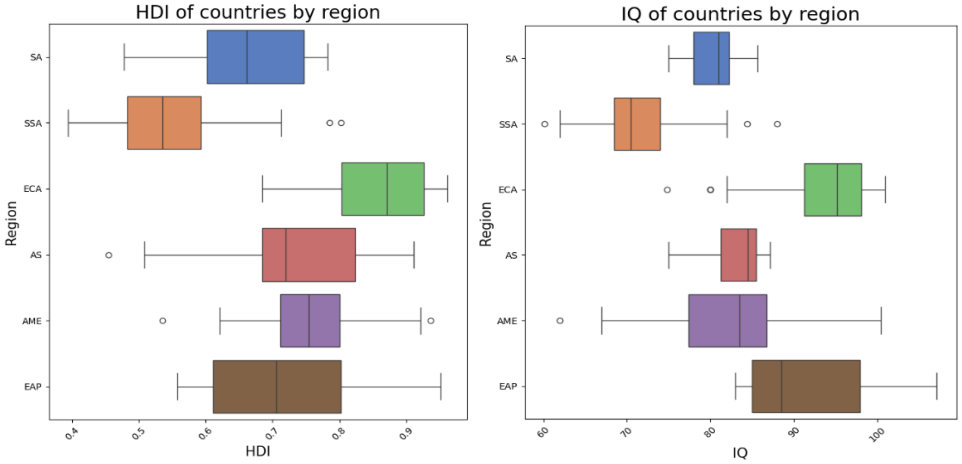

Figure 2

For HDI five outliers were found and for the IQ six were found:

In the Sub-Saharan Africa region:

Two upper outliers, Mauritius and Seychelles with HDIs of 0.802 and 0.782 where found. As compared to the third highest in the area, South Africa with 0.713. Similarly the two countries had 88 and 84.4 IQ scores repsectiely were found, with the next lowest IQ score being that of Madagascar, with an IQ of 82. Malawi was a lower IQ outlier, with a score of 60.1 as compared to the score of 62 from Gambia with the next lowest IQ score

In the Asia South region:

Syria was an upper outlier in the Asia South region, with a HDI of 0.577 compared to the next highest country of Djibouti with a score of 0.509.

In the Americas region:

Haiti was an lower outlier in the Americas region, with a HDI of 0.535 compared to the next lowest value of 0.621 of in Honduras. Saint-Lucia was a lower IQ outlier in the the americas region. with a score of 62, the next score being 67 of

Dominica. Unsuprisingly, as we had made all the Americas into one category, and due to fact that USA and Canada are top preformers in terms of HDI , they are found to be uppers outliers in the Americas region. The USA having a HDI of 0.921 and Canada having 0.936.

In the Europe and Central Asia Region:

Kyrgystan was found to be a lower outlier with a score of 74.8 with the next lowest score being 80.0 of Uzbekistan.

4.2 Bivariate analysis

Since the main variables of intrest in this investigation was IQ and HDI, and specifically the relation between these two, a scatter plot between the two was exectued, as can be seen in Figure 3**.** An important point of interest for us was how these correlations differed between different regions, which was why we opted to use colourcoded regions to get a preliminary understanding of correlation patterns. What can be concluded through observing the graph is that there indeed seems to exist a positive linear correlation between HDI and IQ. The quantitiatve results of the Pearson and Spearman correlations supports this as well. The Pearson correlation is 0.816, meaning that it has fairly strong linear correlation, such that when HDI increases the IQ follows suit. To confirm whether the correlation could be better explained by a monotonic function, a spearman test was also conducted, this yielded a result of 0.824, since this is close to the result of the pearson test, we can conclude that the relationship cannot be better described by a montotonically increasing function, but that a linear relationship acts as the best representation between the variables.

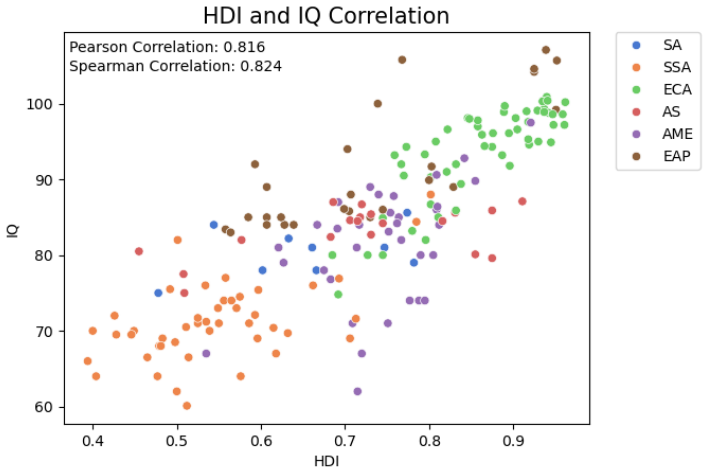

Figure 3

However, visually inspecting the scattering of the points indicates that the relationship between the variables seems to be stronger where both HDI and IQ higher, whilst the points closer to the origin are more scattered. As can be seen in the univariate analysis, there is a large difference between the regions range of values for the HDI and IQ scores. Hence, there seems to be a difference in correlation from region to region. Visually, this shows itself intituively, since the colour coded regions appear to be differing in how large how large their spread is.

To confirm this, indiviudal scatterplots of each region were made along with a barchart to compare the difference in pearson and spearman correlation between the different regions. The results shown in Figure 4 prove that there is a big discrepeancy between the correlation in the regions. Where as East Asia and Pacific and Europe and Central Asia exhibit a similar correlation found in the entire dataset, the other regions had lower correlation coefficents with South Asia only having a Pearson correlation of 0.410.

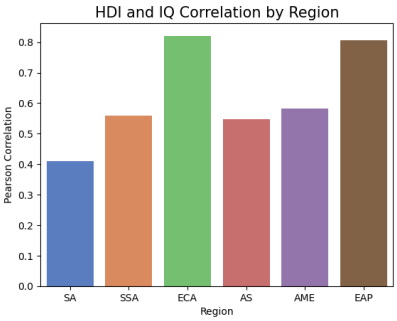

Figure 4

5 DISCUSSION AND PRELIMINARY CONCLUSIONS

The findings from our analysis provide evidence in support our hypotheses concerning HDI and IQ:

1. The average IQ and Human Development Index are indeed positively correlated on a Country Level.
2. Less developed regions are indeed more prone to outliers.
3. The higher correlations between HDI and IQ observed in western regions suggest that the IQ measure of intelligence may indeed be more reflective of the educational and cognitive styles predominant in Western societies, and is therefore and indicator for bias.

The implications of these findings are manifold. For policymakers and educators, the correlation between HDI and IQ underscores the importance development strategies that encompass not only economic growth but also improvements in health, education, and living standards. For researchers, the observed biases and outliers invite further investigation of IQ biases and the development of more culturally sensitive measures of cognitive ability.

In conclusion, this analysis not only supports the hypothesized relationships between Human Development and Analytic Inteligence but also prompts critical reflections on the measures we use to assess these constructs.